Naval Research Laboratory

Washington, DC 20375-5000

NRL Report 9206

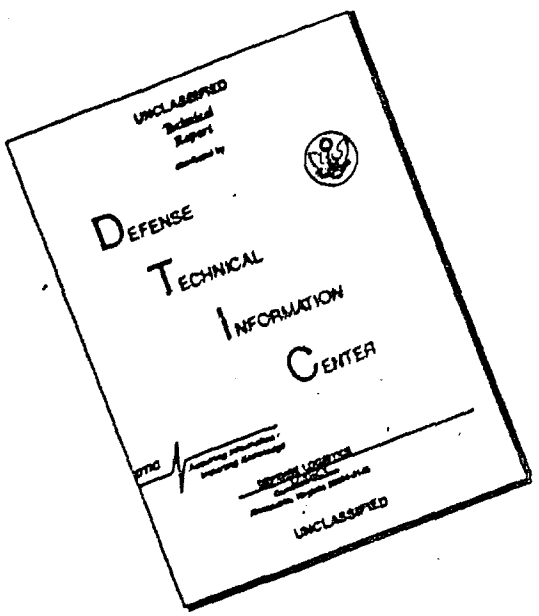
Voice Preprocessor for Digital Voice Applications

G. S. KANG, L. J. FRANSEN, AND T. M. MORAN

Human-Computer Interaction Lab
Information Technology Division

September 11, 1989

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a REPORT SECURITY CLASSIFICATION UNCLASSIFIED	
2a SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited
2b DECLASSIFICATION/DOWNGRADING SCHEDULE	
4 PERFORMING ORGANIZATION REPORT NUMBER(S) NRL Report 9206	5 MONITORING ORGANIZATION REPORT NUMBER(S)
6a NAME OF PERFORMING ORGANIZATION Naval Research Laboratory 6b OFFICE SYMBOL (If applicable) Code 5531	7a NAME OF MONITORING ORGANIZATION
6c ADDRESS (City, State, and ZIP Code) Washington, DC 20375-5000	7b ADDRESS (City, State, and ZIP Code)
8a NAME OF FUNDING/SPONSORING ORGANIZATION Space and Naval Warfare Systems Command 8b OFFICE SYMBOL (If applicable) PMW-151-21	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c ADDRESS (City, State, and ZIP Code) Washington, DC 20360	10 SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO: 61153N PROJECT NO: X7290-CC TASK NO: 55-0114-C-9 WORK UNIT ACCESSION NO: DN 280-290

11 TITLE (Include Security Classification)
Voice Preprocessor for Digital Voice Applications

12 PERSONAL AUTHOR(S)
Kang, G. S., Fransen, L. J., and Moran, T. M.

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
Interim	FROM ____ TO ____	1989 September 11	42

16 SUPPLEMENTARY NOTATION

17 COSATI CODES			18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)	
FIELD	GROUP	SUB-GROUP	> Speech preprocessing	Microphone response equalization
			Speech conditioning	Automatic gain control
			Perception of speech distortion	Digital antialiasing filtering

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

A voice processor operating satisfactorily in laboratory environments with carefully prerecorded speech samples often fails to operate satisfactorily with live speech. Potential reasons are: (1) the speech level may be too high or too low; (2) the speech signal may have too much interference (ambient noise, breath noise, 60 Hz hum, digital noise in analog circuits, a DC bias (caused by component aging, etc.) generated at the analog-to-digital converter output); (3) the microphone frequency may be severely distorted; (4) the speech signal from the existing audio system, in certain operating environments, may be improperly coupled to the front-end circuit; (5) the speaker may be talking too fast or may have an improper mouth-to-microphone distance, or the speech spectra may lack high-frequency energies.

In this report, we have generated a comprehensive design for a speech preprocessor that removes interferences, adaptively equalizes frequency anomalies, and conditions speech for speech encoding, speech recognition, speaker recognition, or extraction of verbal or nonverbal information from speech.

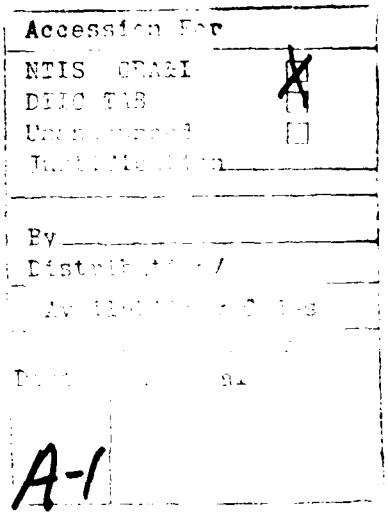
(Continues)

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION
[X] UNCLASSIFIED/UNLIMITED [ ] SAME AS RPT [ ] DTIC USERS	UNCLASSIFIED
22a NAME OF RESPONSIBLE INDIVIDUAL George S. Kang	22b TELEPHONE (Include Area Code) (202) 767-2157 22c OFFICE SYMBOL Code 5531

DD Form 1473, JUN 86

19. ABSTRACT (Continued)

It has taken over a decade of R&D to deploy voice processors. Once they are deployed, they will not be easily replaced for various reasons. Therefore, voice processors must be designed such that they will operate satisfactorily even under unexpected operating environments long after they have been deployed. This report is written with that goal in mind.



CONTENTS

VOICE PREPROCESSOR FOR DIGITAL VOICE APPLICATIONS

INTRODUCTION

For many years, digital voice processors have been used to transmit speech information at low bit rates in secure voice applications. Digital voice processors are increasingly being used for recognizing speech or speakers to facilitate human-computer interaction (Fig. 1). In any of these applications, an indispensable part of the process is the characterization of the speech spectrum. Recently, numerous digital speech processing techniques have been developed for this purpose with the linear predictive coding (LPC) being the most widely used technique. We have also investigated LPC analysis/synthesis for improving low-bit-rate voice encoding [1,2].

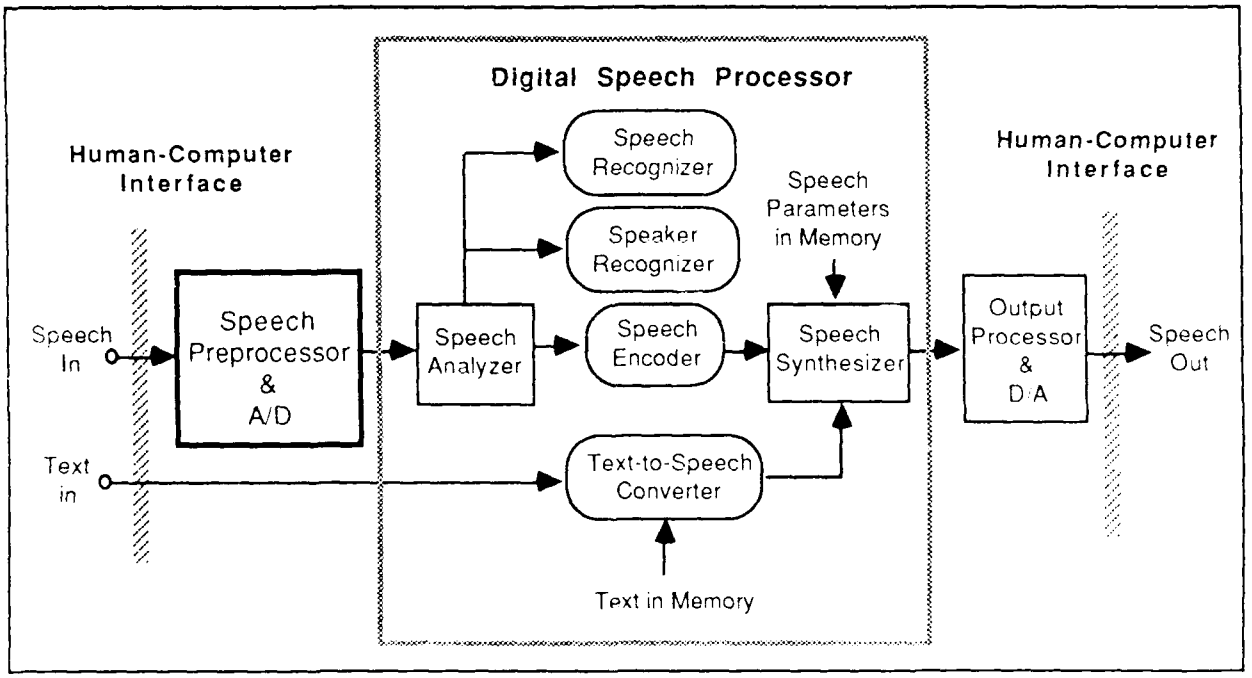


Fig. 1 — Digital voice processors for secure voice and human-computer interactive applications. The speech preprocessor (indicated by a thick-lined box) conditions the speech signal for subsequent speech analysis for various applications. The quality of the speech preprocessor significantly affects the overall system performance. During the early days of 2400-b/s voice encoder development, refinement of the front-end analog circuits alone improved the speech intelligibility by as much as five points, which demonstrates the importance of the input audio processor.

Manuscript approved February 27, 1989

Strangely, no comprehensive investigation has been related to the requirements of a speech preprocessor (often known as the front-end processor or speech I/O), which is the critical link between humans and computers. Over the years, we have designed audio circuits based on somewhat lax performance specifications because severely distorted speech can still provide acceptable speech quality. For example, the severe distortion in the speech waveform from a carbon microphone is due to the random modulation of the electrical resistance caused by the movement of the carbon granules. But the quality of the telephone speech is deemed acceptable to a majority of its users. Because our ears are tolerant to speech distortions, speech input circuits have often been haphazardly designed.

But the speech analyzer in the digital voice processor is not a human ear. A speech waveform anomaly not objectionable to the human ears (viz., peak clipping of the speech waveform) can cause a significant deterioration in the estimated speech spectrum. The speech waveform has a wide dynamic range (60 dB or more), and it is difficult to maintain a correct speech level. An improper speech level is one of the reasons why a voice processor optimized in the laboratory by the use of carefully prerecorded speech often fails in the field because of the varying levels of live speech.

The speech preprocessor presented in this report is more than a conventional front-end speech I/O device comprised of an amplifier, an antialiasing filter, and an analog-to-digital (A/D) converter. Our preprocessor self-adjusts the speech level and equalizes the microphone frequency response and the spectral tilt of voices; it also removes various interferences detrimental to speech analysis, such as distorted microphone response, breath noise from the microphone, digital noise in the analog channel, 60 Hz hum, unintentional DC bias from the A/D converter, and external ambient noise. In other words, the speech preprocessor conditions the speech signal to produce the best speech analysis result for the intended applications (i.e., speech encoding, speech recognition, or speaker recognition).

Note that many preprocessing operations will be digital rather than analog because of the following advantages:

- *Miniaturization*—Elaborate analog circuits are a hindrance to miniaturizing voice processors. Over the years, weight and power consumption of digital voice processors have declined steadily (Fig. 2), and this trend will continue. Our approach to speech preprocessing lends itself to future hardware miniaturization.

- *No Aging Problem*—The performance will not degrade because of aging components in the analog circuits.

- *Flexibility and Power*— Digital processing has more flexibility. For example, filter characteristics of a digital filter can achieve more ideal filtering characteristics (i.e., steeper cutoff rate and linear phase response), and they can be altered more conveniently by changing weights. If needed, filter weights can adaptively be changed based on whether the speech signal is voiced (which needs sharp cutoff) or unvoiced (which does not need sharp cutoff to bring out-of-band speech energies into the passband).

The topics included in this report have come to our attention while working with various voice processors over the past 15 years. We hope that our thoughts and experience will guide the designers of future secure voice and human-computer interactive voice systems.

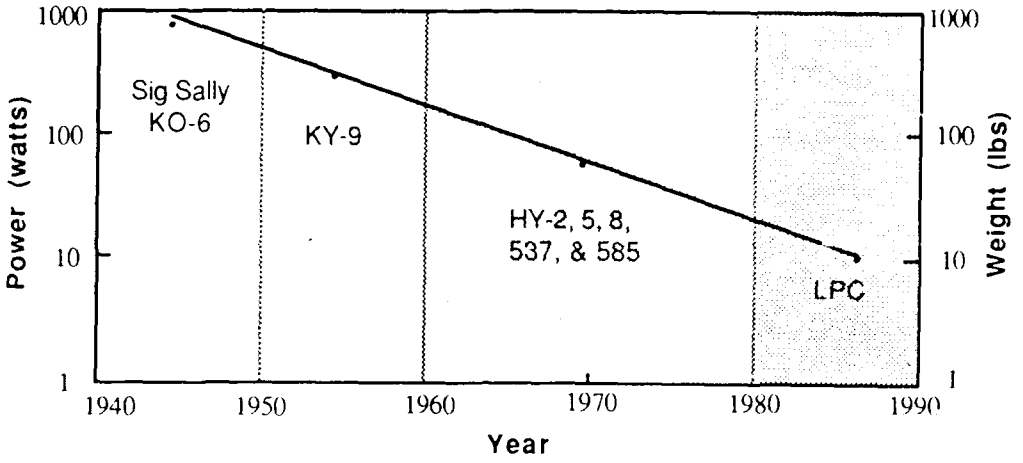


Fig. 2 — Power consumption and weight of low-bit-rate digital voice processors. This figure shows how the advancement of digital component technology has contributed to the reduction of both weight and power consumption. Sig Sally was packaged in a dozen 6 ft racks; KY-9 was contained in seven 19-in. racks (weighing 500 lb). In our preprocessor, the operations heretofore carried out by analog circuits are relegated to digital signal processing. Reduction of analog processing has been an influential factor in the miniaturization of voice processors.

BACKGROUND DISCUSSIONS

Wide Dynamic Range

Speech is a difficult signal to interface with a signal processor because speech has a wide dynamic range. Peak amplitudes of vowels are often 40 dB greater than peak amplitudes of fricatives (Fig. 3). In addition, a 20 dB difference in loudness exists from one speaker to another. Therefore, a front-end processor must have a dynamic range of at least 60 dB. Otherwise, vowel waveforms will be clipped frequently, or weak fricatives will be lost. If any of these occur, the performance of the voice processor will be degraded.

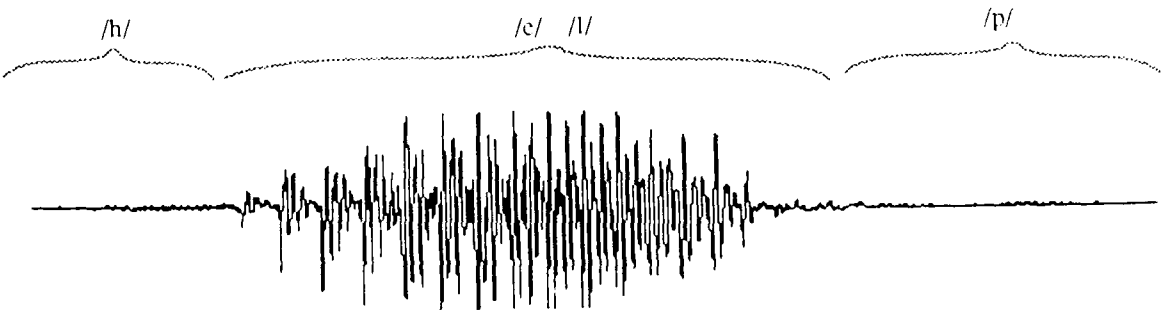


Fig. 3 — Speech waveform of "help." The peak amplitude of speech varies as much as 40 dB within a fraction of a second, as noted from the fricative /h/ to the vowel /e/ in this example. A wide dynamic range is a notable characteristic of the speech waveform. An improper front-end gain will distort the speech waveform. One critical function of the processor is to maintain a proper input speech level.

Perceptual Tolerance to Speech Distortion

But distorted speech is not too objectionable to the human ear. It has been long known that so-called "peak-clipping" of speech has little perceptual effect on intelligibility and negligible effect on quality, even if 10 to 12 dB of the highest voice peaks are eliminated. During World War II, the U.S. Army Signal Corps Engineering Laboratories tasked the Psycho-Acoustic Laboratory of Harvard University to investigate the maximum degree of amplitude distortion tolerable in a communication system (i.e., analog communication systems such as the telephone). It was found that if speech were differentiated prior to clipping, even infinite clipping retained 90% to 95% of the original intelligibility of nonsense monosyllables [3]. For two reasons, human ears are insensitive to amplitude distortions:

- *Harmonic Structure of Voiced Speech Spectrum*—Voiced speech (vowel sounds) is generated by periodic ringing of the vocal tract by the glottis. Therefore, the voiced speech waveform is periodic at the pitch rate (Fig. 4); its spectrum is concentrated at pitch harmonics (Fig. 4). Note that amplitude distortions of voiced speech do not create cross products of frequencies that fall between pitch harmonics (or inharmonic sounds). Hence distorted speech is not too objectionable to our ears.

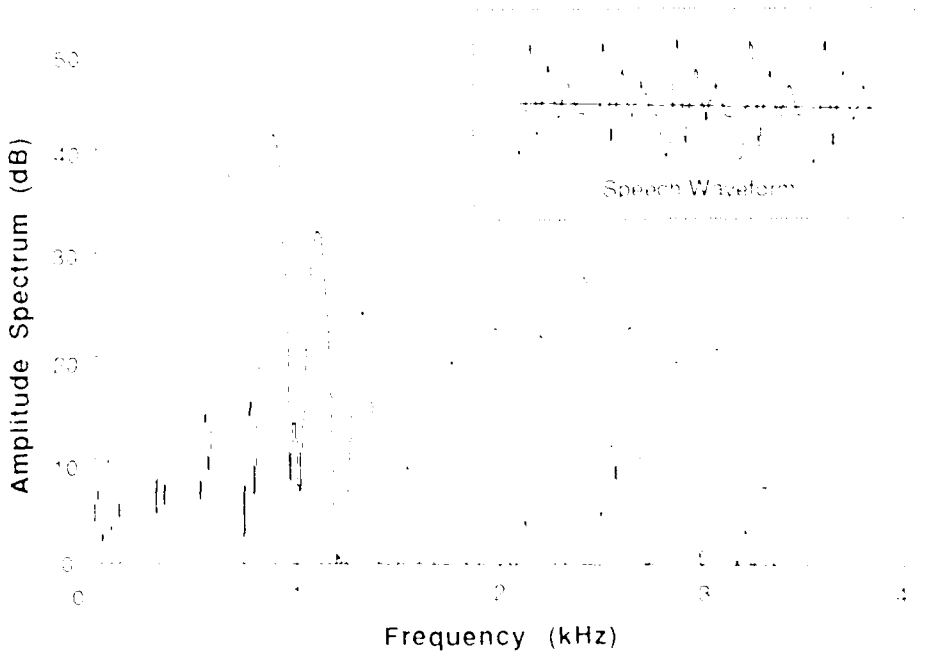


Fig. 4. Speech waveform of vowel /æ/ and its frequency spectrum. Because the speech waveform is repetitive at the pitch rate, its frequency components are concentrated at pitch harmonics. Thus, cross products of frequencies generated by nonlinear distortions are also concentrated at pitch harmonics. That is why distorted speech from carbon microphones is not too objectionable to human ears, whereas music (whose spectrum is entirely different from the voiced speech spectrum) sounds terrible over the telephone.

- *Variability of Unvoiced Speech Spectrum*—Unvoiced speech is created by turbulent air through a constriction somewhere in the vocal tract. Since the time waveform is random, its spectrum is also random (Fig. 5). For a given unvoiced speech, its spectrum varies widely from speaker to speaker because each has different lip, tongue, and teeth clearances. Distorted unvoiced speech of one speaker can sound like undistorted unvoiced speech of another person. This is why we do not perceive distorted unvoiced speech as being objectionable.

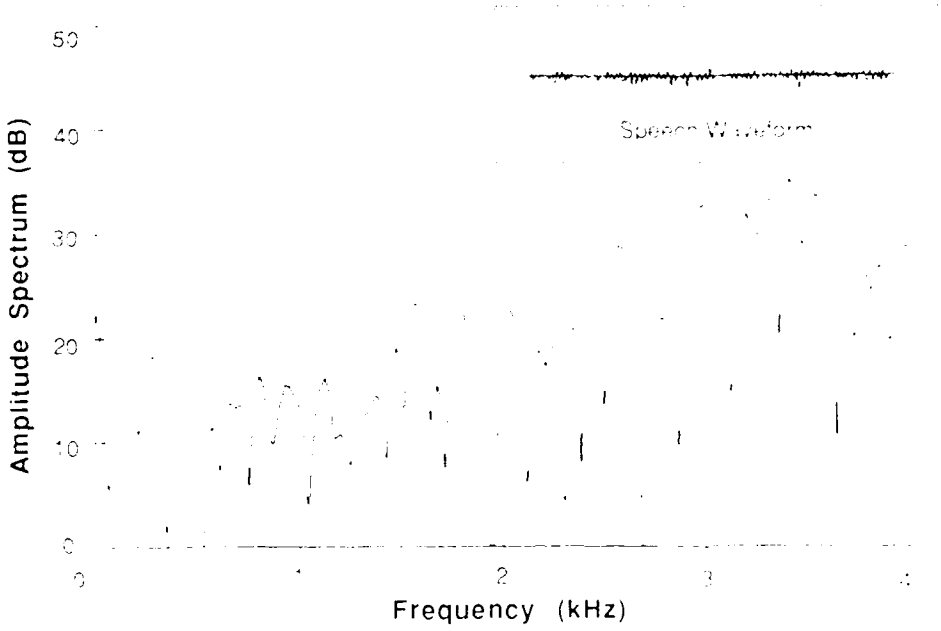


Fig. 5 — Speech waveform of unvoiced speech "s" and its frequency spectrum. Unlike the voiced speech spectrum, the unvoiced speech spectrum is random. A distorted unvoiced spectrum of one person may be similar to an undistorted unvoiced spectrum of another person. That is why distortions of unvoiced speech are not objectionable to our ears.

Perceptual Tolerance to Stationary Phase Shift

In addition, our perception is insensitive to certain kinds of phase distortions. For example, a stationary phase shift of the speech spectrum is not discernible to us. To illustrate this phenomenon, the speech waveform is passed through an all-pass filter whose amplitude response is flat,

$$A(f) = 1, \tag{1}$$

and phase response is a quadratic function of frequency (i.e., the group delay is a linear function of frequency),

$$\phi(f) = 6\pi(f/4000)^2 \text{ radians}, \tag{2}$$

where f is in hertz, and half the sampling frequency is 4000 Hz. Although the input and output speech waveforms look different (Fig. 6), they sound exactly alike. They must be heard to be believed!

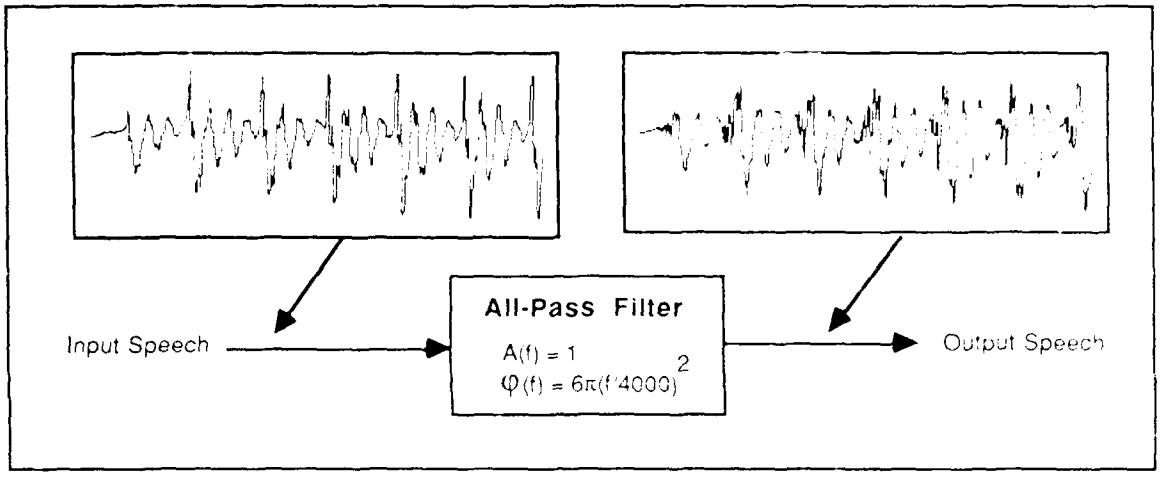


Fig. 6 — Input and output speech waveforms for the all-pass filter. As noted, the output speech waveform is distorted, but both the input and output speech waveforms sound exactly alike. Our hearing is blind to time-invariant phase shift.

Effects of Speech Distortion on Spectral Estimation

It is significant to point out, however, that what is acceptable to human perception is very different from what is acceptable to the digital speech processor, which tries to estimate the speech spectrum by a limited number of parameters. For example, peak clipping of speech is highly detrimental to speech analysis, such as the LPC analysis.

The LPC analysis is based on the assumption that a given speech sample x_t is predicted by a weighted sum of past samples,

$$x_t = \sum_{k=1}^{10} a_k x_{t-k} + \epsilon_t, \tag{3}$$

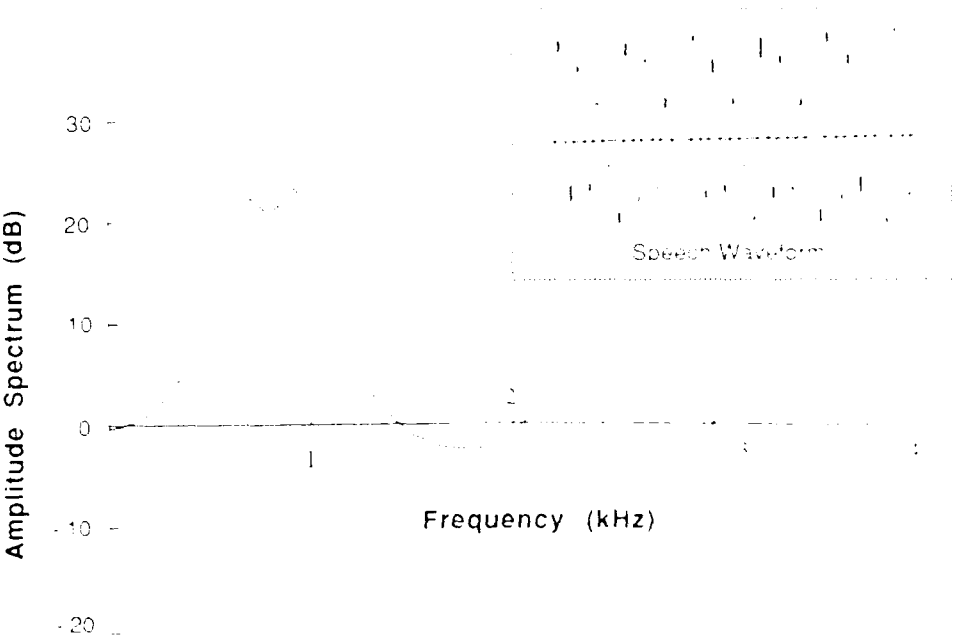
in which a set of weighting factors a_k is estimated by minimizing the mean-square prediction errors. The prediction principle does not hold well when the speech waveform is clipped. As a result, the estimated speech spectrum becomes erroneous. Figure 7 illustrates the effect of waveform clipping on the LPC spectrum.

CRITICAL DESIGN ISSUES

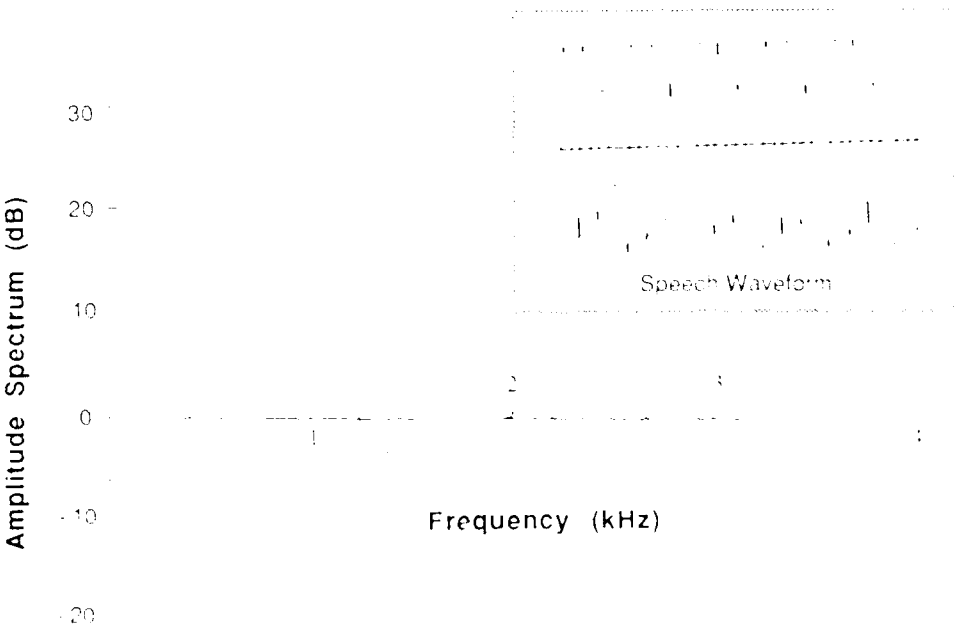
The critical issues related to the design of a voice preprocessor include microphone frequency response equalization, input coupling, automatic gain adjustment, digital implementation of the antialiasing filter, and reduction of various forms of interference (Fig. 8). Each item is discussed in a subsequent section.

1. Microphone

Noise-cancelling microphones are used in all military platforms. The noise-cancelling microphone attenuates undesirable background acoustic noise, and it also attenuates unintentional information (including human voices in the background) leaking into the microphone.



(a) Without speech waveform clipping



(b) With speech waveform clipping

Fig. 7 — Effect of speech waveform clipping on estimated speech spectrum by LPC analysis. As illustrated, a small amount of clipping (11%) alters the two upper formant frequencies by approximately 10 dB

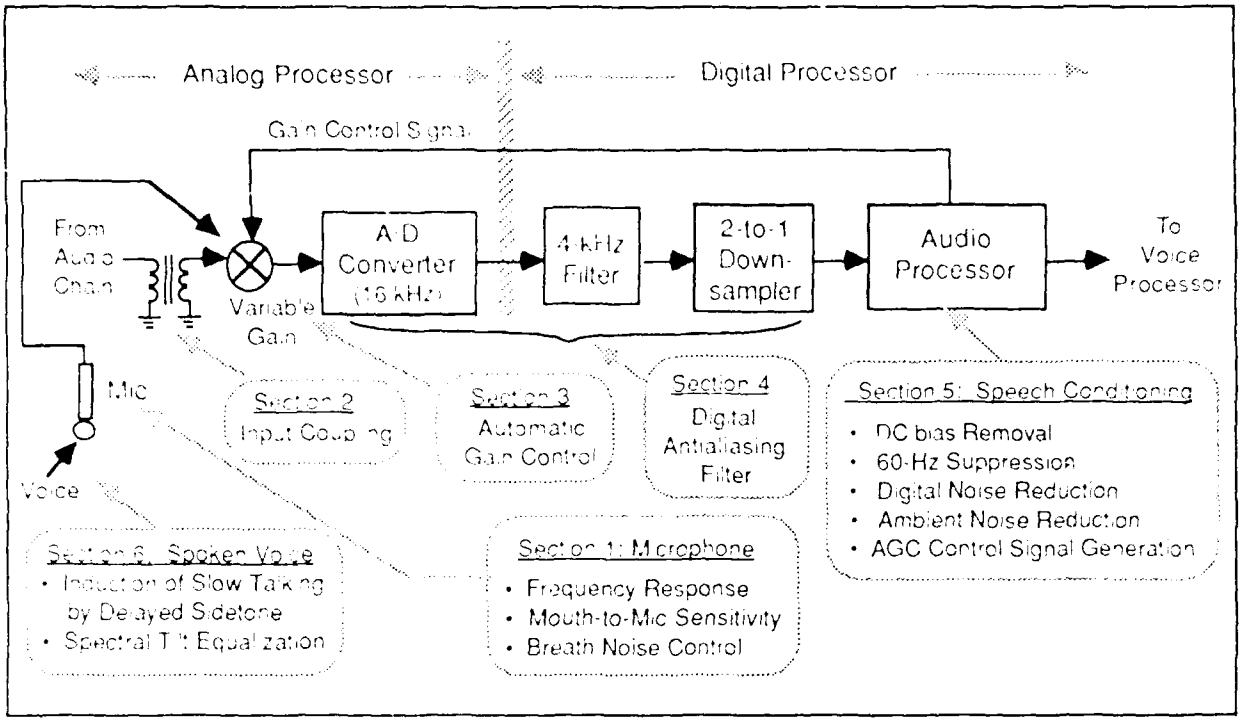


Fig. 8 A preprocessor for voice processing applications. The preprocessor automatically adjusts the speech level, removes speech interference, and digitizes the speech signal with a nearly ideal antialiasing filter. The topics of discussions are indicated in shaded boxes.

The noise-cancelling microphones currently in use are the first-order gradient microphones that were developed in the 1930s by Harry Olson [3]. The output of this type of noise-cancelling microphone is proportional to the pressure difference between two closely spaced elements. The output of a noise-cancelling microphone caused by a sinusoidal source is expressed by

$$\Delta P = P_m \frac{\sin(ct - r)}{r^2} + \frac{2\pi}{\lambda} \frac{\cos(2\pi\lambda)(ct - r)}{r} D \cos\theta, \tag{4}$$

where r is the distance to the sound source, P_m is a proportional constant, D is the separation of the two microphone elements, λ is the waveform length of the sound source, c is the speed of sound, and θ is the signal arrival angle measured from the axial direction [3].

The near-field response (which has r^2 in the denominator in Eq. (4)) is for speech, and it is independent of frequency. The far-field response (which has r in the denominator) is for noise, and it is directly proportional to frequency; the frequency response has a -6 dB/octave attenuation characteristic toward low frequencies. This ideal frequency response, however, is seldom attained in practice because of the complex mechanical structure around microphone elements, and no two microphones from differing manufacturers have similar frequency responses.

In certain operating environments, the speech signal may come from existing intercommunication systems or audio chains. In this case, voice processors (either voice encoders, speech recognizers, or speaker recognizers) must work with the existing microphone. Thus it is worthwhile to review some of the better known noise-cancelling microphones to assess the amount of frequency equalization

we need to equalize each microphone. We show typical mouth-to-microphone sensitivities to indicate the degree of speech level fluctuations expected if the microphone is not held properly. We point out also that the puff screen is essential for noise-cancelling microphones because they tend to distort the onsets of plosives.

1.1 Frequency Response Equalization

NRL has surveyed the existing microphones, audio systems, and ambient noise characteristics in various tactical platforms [4]. This program was tailored to assess how the Advanced Narrowband Digital Voice Terminal (ANDVT) would perform with acoustic noise at the input and stress, vibration, and acceleration applied to the talker. To do this, both trained and "walk on" speakers read Diagnostic Rhyme Test (DRT) words and Diagnostic Acceptability Measures (DAM) sentences from military platforms while engaging in realistic maneuvers. Through this program, a vast amount of data related to microphones and audio systems was collected.

Most of the presently deployed noise-cancelling microphones were originally designed for analog voice communication systems where a flat frequency response was not essential. In digital voice processors, however, the presence of microphone response peaks affects adversely the estimation of the speech spectrum. Not all presently deployed noise-cancelling microphones have a flat frequency response, as shown in Fig. 9. They must be equalized to have a flat response.

Microphones for tracked vehicles (e.g., the M-87 and M-138) were originally designed to attenuate low frequencies to filter out mechanical rumbles. For speech analysis, however, a lack of low frequencies is detrimental: (a) pitch tracking and voicing decision become less reliable, and (b) recognition of nasals (/m/, /n/ or /ng/) become difficult because they have mainly low-frequency components. Thus our recommendations are:

- The frequency response should be restored to the ideal flat response between 150 to 3800 Hz.
- A more effective digital preprocessing should be used to eliminate noise.

In Section 5 we present a noise suppression method that equalizes microphone frequency response. In this method, the noise suppression is carried in the frequency domain. Therefore, the microphone response equalization can be effected conveniently as an integral part of the noise suppression at a small computational cost.

1.2 Mouth-to-Microphone Sensitivity

The induced speech level of a noise-cancelling microphone is significantly dependent on the mouth-to-microphone distance. An improperly held microphone is a major cause of speech-level fluctuations that are highly detrimental to speech analysis. Figure 10 shows the amount of speech attenuation expected when the microphone is moved 1.4 to 1 in. from the mouth. The average magnitude of speech attenuation is somewhere around 12 dB, which is rather significant, considering that the microphone could be easily moved by 3-4 in., even while trying to hold the microphone steadily. Our recommendation is that the preprocessor shall have a self-adjusting amplifier gain. In Section 3 we present an effective software-controlled automatic gain control (AGC) mechanism.

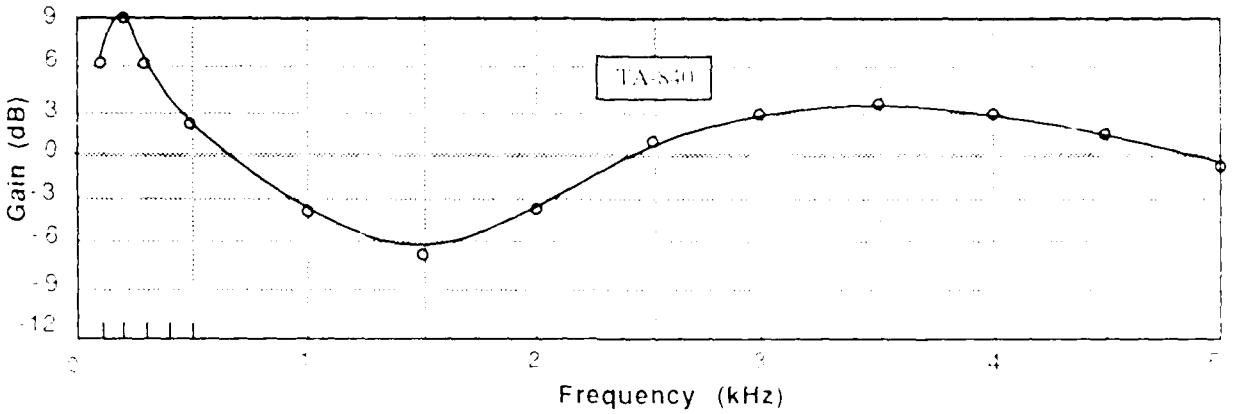


Fig. 9(a) — Frequency response of the TA-840. The TA-840 handset is widely deployed for naval communication in shipboard environments. Several different noise-cancelling microphone elements have been developed for the TA-840. This particular microphone element produces bass-heavy and not-too-intelligible speech sounds. In addition, bass-boosted speech can generate acoustic feedback between the microphone and intercom speaker. As noted, the frequency response is far from flat, it should be equalized before performing digital speech processing. The method of equalizing the frequency response is given in Section 5.

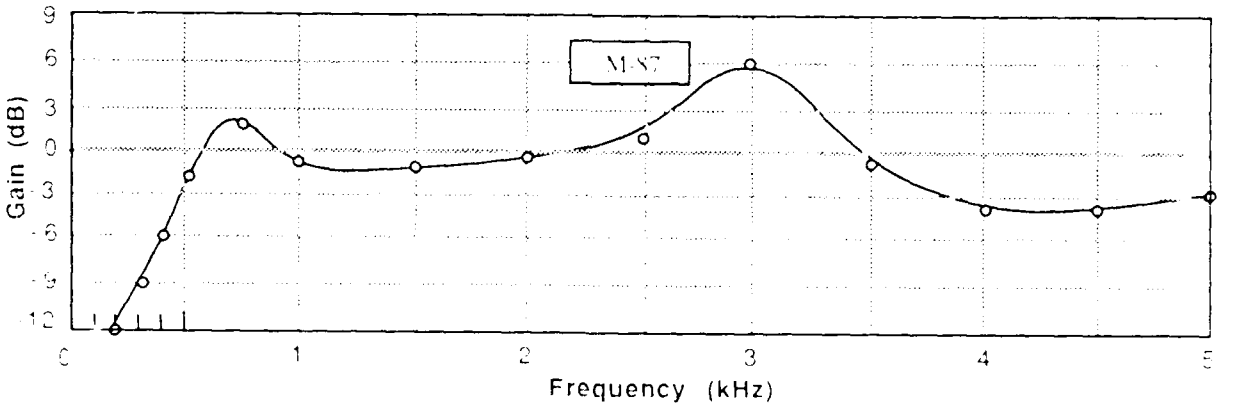


Fig. 9(b) — Frequency response of the M-87 noise-cancelling microphone. The M-87 is a boom microphone that has been widely used by the Navy and Air Force in airborne and tracked vehicles. Low frequencies are severely rolled off (a −3 dB gain at approximately 500 Hz) to attenuate mechanical rumbling. According to tests, the M-87 outperformed the TA-840 when used as an LPC front-end microphone. As in the TA-840, the frequency response of the M-87 should be equalized prior to speech processing.

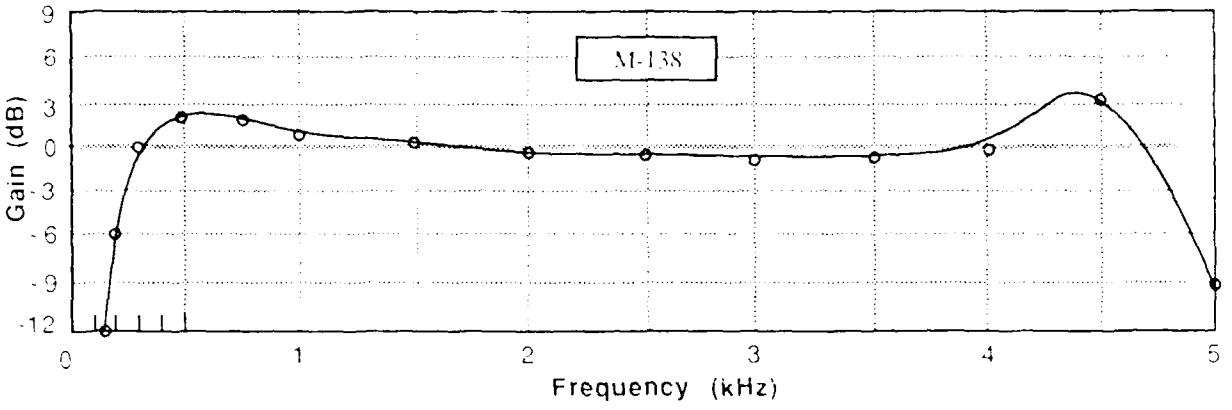


Fig. 9(c) – Frequency response of the M-138 noise-cancelling microphone. The M-138 is a boom microphone and has been used interchangeably with the M-87. Since the speech waveform will be filtered at 4 kHz, a peak around 4.4 kHz is inconsequential to the performance of the voice processor. The M-138 has a better frequency response than either the M-87 or TA-840. But it did not work as well as the M-87 at tank platforms where low-frequency noise is predominant. The reason is that the M-87 severely attenuated low frequencies, whereas the M-138 microphone actually boosts low frequencies that should be equalized.

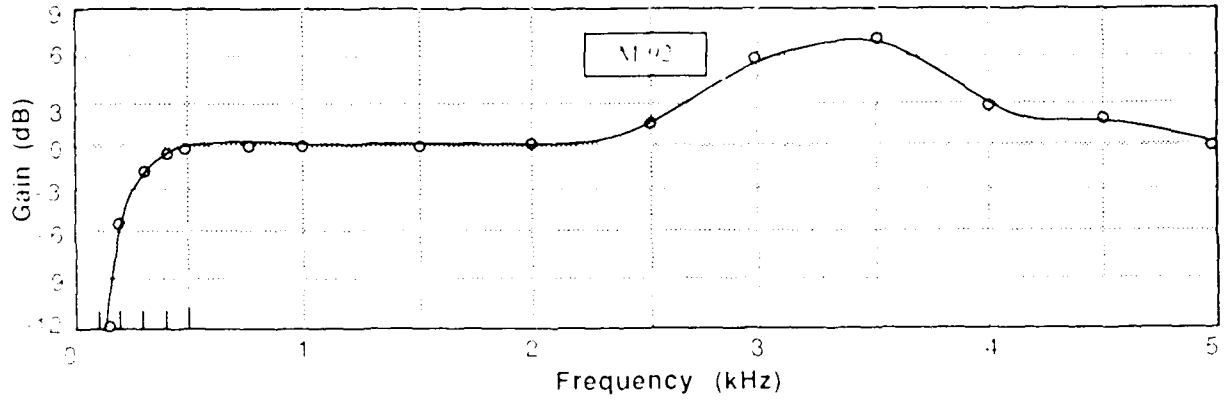


Fig. 9(d) – Frequency response of the M-92 noise-cancelling microphone. The M-92 is a handheld microphone often used in the P3C platform. The frequency response below 2 kHz is nearly ideal. A 7 dB peak near 3.5 kHz should be equalized. According to tests, the M-92 was one of the better microphones for the LPC front end.

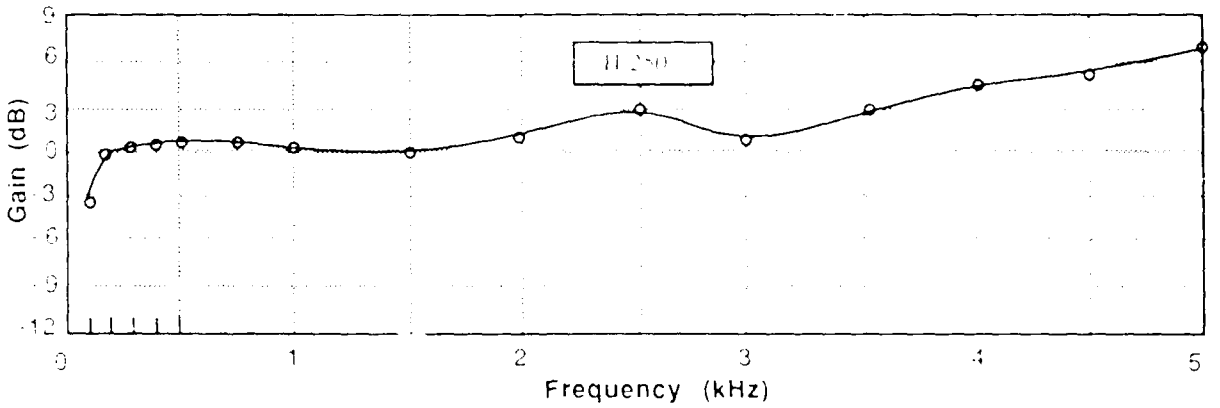


Fig. 9(c). Frequency response of the H-250 handset. The H-250 is a handheld, noise cancelling microphone used in conjunction with the jeep radio. The H-250 is a far better microphone than the preceding microphones.

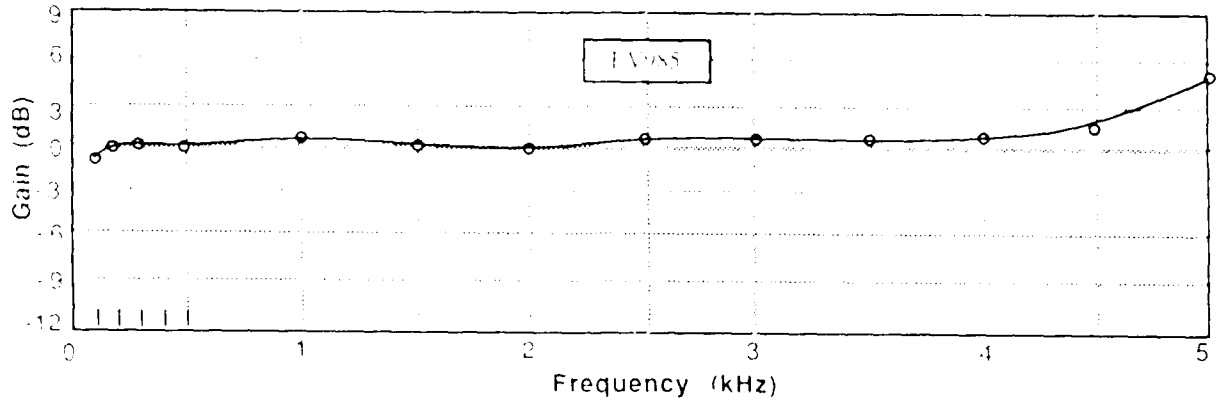


Fig. 9(d). Frequency response of the EV-985 noise-cancelling microphone. This electret boom microphone was developed for the Army. This is the best noise-cancelling microphone we have tested to date. No frequency equalization is needed.

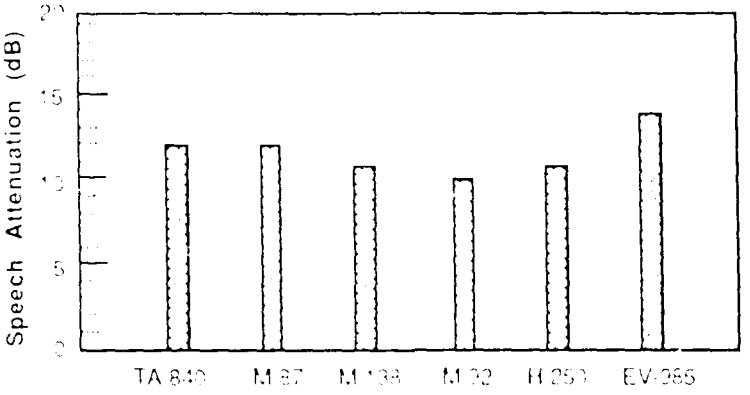
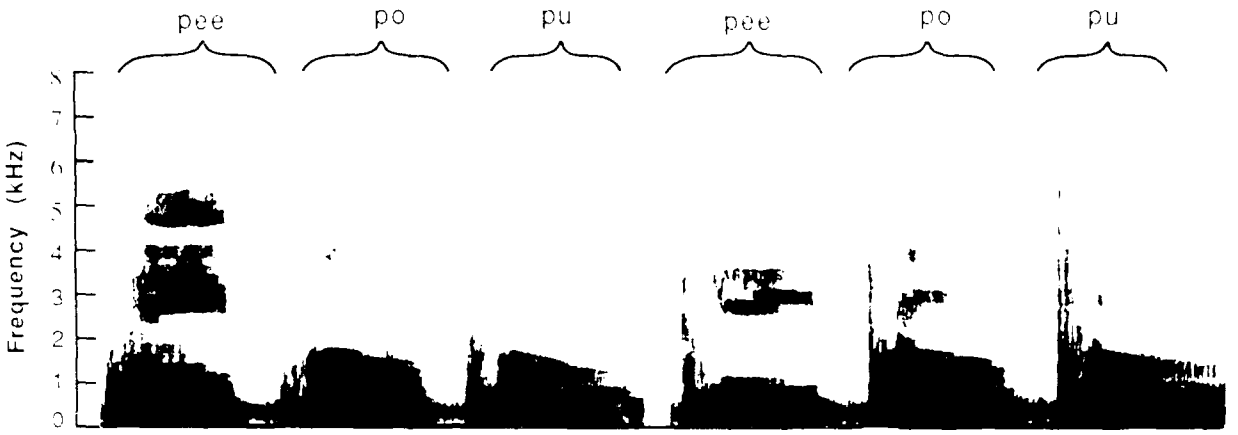


Fig. 10. The loss of microphone output level because of microphone misplacement. When the microphone is moved from an optimum distance of 1.4 to 1 in. in the axial direction, the speech level is decreased by the amount indicated in this figure.

1.3 Breath Noise

The leading edge of a sound wave is often spattered on impact at the surface of the microphone, creating a burst of noise at the speech onset, particularly at onsets of plosives such as /p/. This phenomenon is more pronounced with the noise-cancelling microphone because, as discussed in connection with Eq. (3), the far-field frequency is similar to a high-pass filter (i.e., a differentiator). The spectrum of /p/ normally has predominantly low-frequency components, but it spreads noticeably after spattering (Fig. 11). The resultant spectrum resembles the spectrum of /t/.



(a) Microphone with shield (b) Microphone without shield

Fig. 11 — Onset spectra of /p/ with and without puff screen. Plosive sounds without a puff screen in the microphone tend to sound like /t/.

The narrowband LPC tends to accentuate these distorted plosive sounds. They are frequently heard as pops, and the voicing decision tends to be "voiced" rather than "unvoiced" as it should be. Not only is the intelligibility of the plosive sound itself reduced, but conflicting burst and vowel transition information (such as a /t/-like burst followed by formant movements typical of /p/) can be confusing to the listener. For the same reasons, the speech recognizer will be confused.

We recommend that a puff screen be used in all noise-cancelling microphones. According to our measurement, the use of a puff screen does not alter frequency response characteristics, although the speech level could be reduced by 1 or 2 dB across the entire passband.

2. Input Coupling

Often, the audio is routed from a subscriber terminal in a communication center through a cable to the voice processor, often through a switchboard. These circuits typically are balanced 600-Ω audio lines (usually grounded center tap) although in some installations they may be unbalanced (one side grounded). Therefore, the I/O should be designed to satisfy both cases to prevent hum pickup, low level, and loss of low frequencies.

To avoid improper input coupling, many systems have a floating input provided by an input transformer (far more protective from transients and RF than a differential input operational amplifier) that works equally well for both balanced and unbalanced inputs. If the transformer coupling is used, there is no need to further attenuate low frequencies by the antialiasing filter because the transformer inherently attenuates low frequencies.

An important specification of the input transformer is the low-frequency cutoff because the transformer size is more or less determined by the lowest frequency to be transmitted at maximum level. We recommend a low-frequency cutoff of 150 Hz. A number of different off-the-shelf transformers are marketed for use with multitone MODEMS and are acceptable for voice applications.

3. Automatic Gain Control

A most difficult requirement for the audio input processor is to maintain a proper speech level prior to the digital voice processor. As discussed in the Background section, a small amount of peak clipping of the speech waveform, caused by a mismatched gain, can cause serious consequences to the estimated speech spectrum.

Good reasons for using a reliable gain control mechanism for voice processors operating in tactical environments are:

- *Improper handling of microphone*—In tactical platforms, noise-cancelling microphones are routinely used to reduce background noise. As discussed previously, the speech level of a noise-cancelling microphone is highly dependent on the mouth-to-microphone distance. An optimum mouth-to-microphone distance is 1/4 in., but when it is increased to only 1 in. by careless handling, the speech level decreases anywhere from 11 to 14 dB (Fig. 10).

- *Shouting*—In military environments, shouting is not unusual because of excessive background noise or tense operating conditions. The speech level easily jumps 10 to 20 dB by shouting.

- *Operating with existing audio systems*—In certain operation environments, the voice processor may be connected to existing intercom or audio systems. They may have different gain levels from one platform to another. An external gain mismatch could be a serious problem for achieving a proper input level.

The audio front end could be equipped with a manual gain control to compensate for the external gain mismatch. Manual gain controls have reportedly not worked well, however, because the operators in the field often did not know how to adjust them properly. Thus it is desirable to have an AGC at the front end to self-adjust the gain in accordance with the input speech level.

Not all AGC devices, however, are suitable for digital voice processors. For example, a fast-attack-and-slow-release AGC is not appropriate for the frame-by-frame spectral analysis used in the digital voice processor. Amplitude variations within the analysis frame caused by the AGC induce errors in the estimated spectrum.

3.1 Recommended AGC

In our approach, the necessary gain is estimated by digital computations. The estimated gain is then fed back to the analog amplifier of the front end of the digital voice processor. The gain is incremented or decremented depending on the error signal—the difference between the reference level and the quantity derived from the speech (which will be defined shortly). We update the gain during unvoiced or silent periods only. In this way the speech amplitude is not altered during voiced speech segments. Figure 12 shows the gain estimation processor.

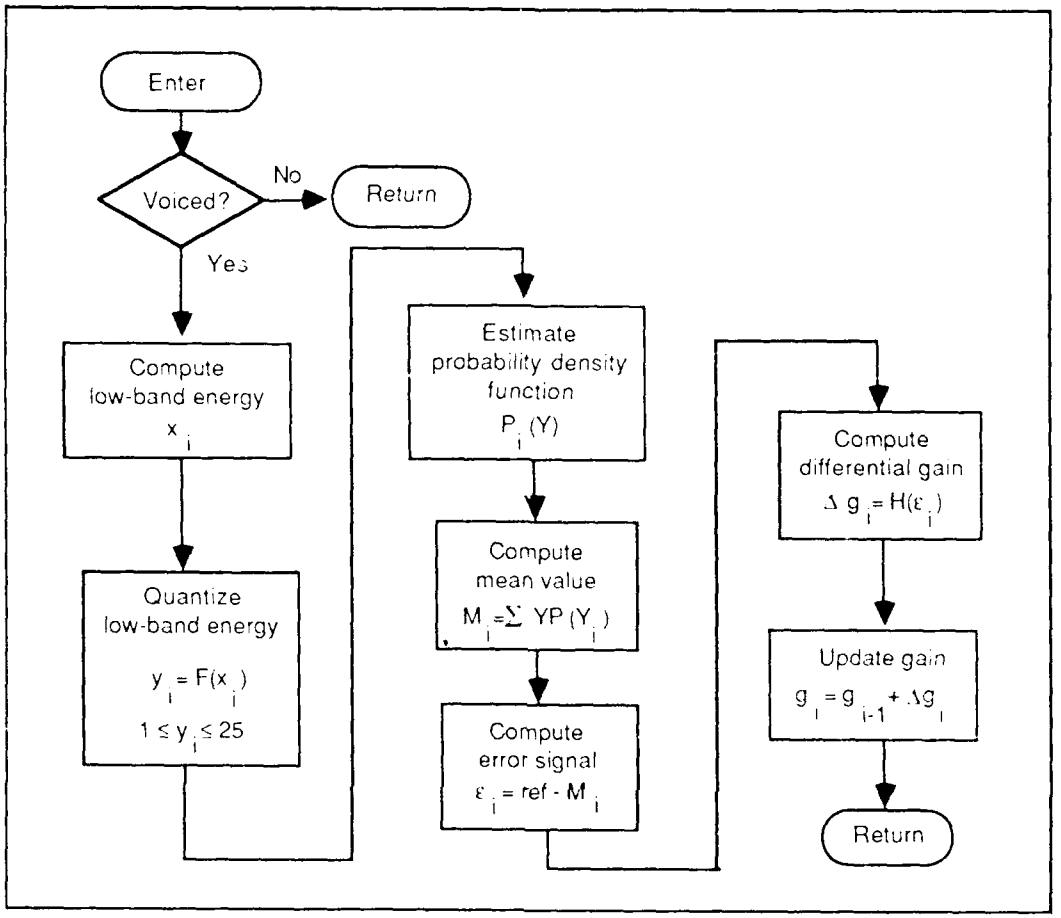


Fig. 12 — Software-controlled AGC. The low-band energy is processed from voiced speech (whose amplitude is as much as 40 dB greater than unvoiced speech). The first-moment of the lowband energy (M_i) is compared with the reference level. The reference level is so chosen that when M_i equals this level, there will be no amplitude clipping. The front-end gain (g_i) is updated during unvoiced or silent periods.

The choice of input variable is critical to the AGC performance. We chose the low-band energy contained below 1 kHz as the input variable (i.e., the first formant amplitude) because it is relatively independent of the nature of speech. Although low-band energy is being averaged over a short time period (20 ms), it has some fluctuations caused by leakage of higher formant frequency components and/or acoustic background noise. Thus we further smoothed the low-band energy through statistical averaging. Time averaging, while simpler, is not as good as ensemble averaging because the amplitude of low-band energy is not uniformly distributed (if so, time averaging would be equivalent to statistical averaging).

To facilitate computations, the incoming low-band speech energy is quantized to one of 25 values, approximately ±20 dB around the reference level (Table 1). We chose a fixed step size of 1.75 dB because we can discern a loudness change of that magnitude. Thus the quantized low-band speech energy is

$$y_i = F(x_i), \tag{5}$$

where x_i and y_i are low-band speech energies before and after quantization respectively, and $F(\cdot)$ denotes the quantization rule listed in Table 1.

Table 1 — Quantization of Speech Low-Band Energy Based on 12-bit Representation of the Speech Waveform. The quantization step size is 1.75 dB. The reference level is 250 or Step 13.

Low-Band Speech Energy (x_i)	Quantized Low-Band Energy (y_i)	Low-Band Speech Energy (x_i)	Quantized Low-Band Energy (y_i)
22 or less	1	306	14
27	2	374	15
33	3	458	16
41	4	560	17
50	5	685	18
61	6	837	19
75	7	1024	20
92	8	1253	21
112	9	1534	22
137	10	1870	23
167	11	2298	24
205	12	2813 or more	25
250 (Reference)	13		

To compute the probability density function of the quantized low-band energy, one register is assigned to each quantization level. When the quantized low-band energy is equal to a particular counter index, the content of that register is incremented by one. The contents of all registers are then short-term averaged by a single-pole filter with a feedback constant of 1/32. Thus,

$$C_i(Y) = C_{i-1}(Y) + [A_i(Y) - C_{i-1}(Y)]/32, \tag{6}$$

where $C_i(Y)$ is the content of the register associated with quantization level during the ith voiced frame. The incremental content $A_i(Y)$ is expressed by

$$A_i(Y) = \begin{cases} 1, & \text{if } y_i = Y \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

In Eq. (6), the feedback constant of 1/32 defines the width of an exponentially decaying window for the register. According to tests with a real-time simulator using a variety of speech samples (noisy as well as quiet), a feedback constant of 1/32 is a suitable choice in terms of fast gain settling without introducing undesirable hunting in the steady state. Note that the feedback constant 1/32 does not directly control the gain update rate; the gain update rate is an incremental gain Δg_i, which appears in Eq. (11).

With updated register counts, the probability density function of the low-band speech energy is computed by

$$P_t Y = \frac{C_t(Y)}{\sum_{Y=1}^{25} C_t(Y)}. \tag{8}$$

The error is defined as the difference between the reference level (REF) and the mean value of the low-band energy. Thus,

$$\epsilon_t = \mathrm{REF} - \sum_{Y=1}^{25} Y P_t(Y). \tag{9}$$

As indicated in Table 1, the reference level is 13 (which corresponds to a low-band energy level of 250 units in a 12-bit A/D conversion). When the mean value of low-band energy equals the reference level, there is no amplitude clipping of any vowels.

The front-end analog amplifier gain in decibels, as denoted by g_t, is incrementally adjusted by

$$g_t = g_{t-1} + \Delta g_t, \tag{10}$$

where the incremental gain Δg_t in decibels is nonlinearly related to the error:

$$\Delta g_t = \begin{cases} 0, & \text{if } |\epsilon_t| \le 2, \\ \dfrac{\epsilon_t + 2}{32}, & \text{if } \epsilon_t < -2, \\ \dfrac{-\epsilon_t + 2}{32}, & \text{if } \epsilon_t > 2. \end{cases} \tag{11}$$

The transform characteristic has a dead zone near the reference level and is linear elsewhere. Thus, if the estimated mean of the low-band energy is within two quantization levels (3.5 dB) of the reference level, no gain adjustment is made. There is a broad range of acceptable update factors. We chose the factor 1/32 after experimenting with various types of speech input, including noisy speech and lengthy two-way casual conversations over a real-time processor. Our decision was based on both transient and steady-state performances, in particular on the gain settling time from an initial gain mismatch as large as −28 dB.

3.2 Prototype Performance

The AGC function described in Section 3.1 has been tested in the NRL-owned programmable voice processor and achieved the following results.

- The AGC established the necessary gain based on past speech statistics. Therefore no additional frame delay was introduced.
- With the use of assembly language, the computation time was 0.55 ms for voiced frames and 0.015 ms for unvoiced frames (one frame is 22.5 ms or 180 speech-sampling time intervals).

- With an external gain mismatch from 0 to −28 dB, intelligibility was virtually unaltered.
- When the input gain was mismatched by as much as −28 dB initially, the steady-state gain was reached within 2 s after the initial onset of voiced speech.
- Once the steady state was reached, there was no noticeable hunting. This condition was based on a 30-min recording of two-way conversations of various speakers.
- No gain pumping was observed in the presence of severe background noise (helicopter noise).

This AGC unit was field tested by using ANDVT over HF channels (Fig. 13). The received speech was recorded at 500 mi away. Transcriptions of recorded voice indicate that the AGC worked satisfactorily for various voices casually speaking in conversational and text-reading modes. Figure 14 is a segment of the speech spectrum of a live message recorded at the receiver.

4. Analog-to-Digital Conversion

In the conventional front-end processor, the analog speech signal is passed through an analog antialiasing filter that sharply attenuates frequencies above 4 kHz (Fig. 15), and the filtered output is digitized at a rate of 8 kHz. In our approach, however, the speech signal is sampled at a rate of 16 kHz, and the necessary filtering is carried out digitally (Fig. 15). According to our experience, there is no need for an 8 kHz low-pass filter prior to A/D conversion because no significant speech energy exists beyond 8 kHz.

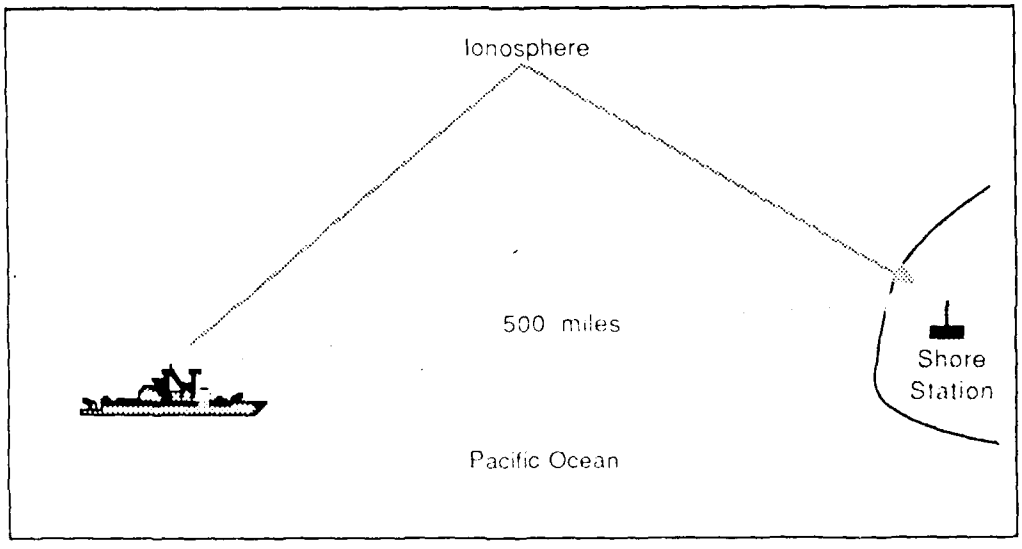


Fig. 13 — HF test of AGC with a secure voice terminal. The AGC was installed in the ANDVT. The 2400-b/s speech was transmitted over the upper sideband of an HF channel from a U.S. Navy ship to a shore station 500 mi away. At the same time, another voice terminal with a manual gain control transmitted the same voice over the lower sideband of the same HF channel. Transcriptions of both speeches at the receiver indicate that the ANDVT with the AGC provided a consistently better matched speech level than the voice terminal with a manual gain control. This is an example of where a manual control is not useful because the operator in the field does not know how to adjust it properly. The use of an AGC at the preprocessor is recommended.

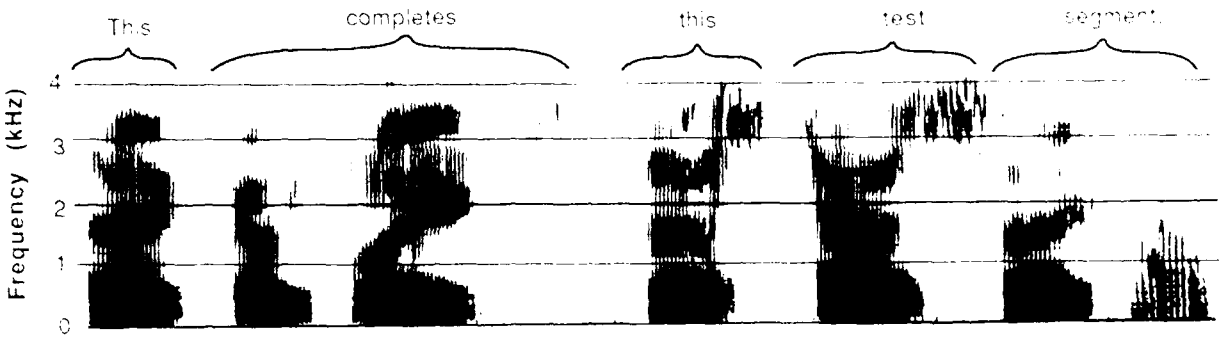
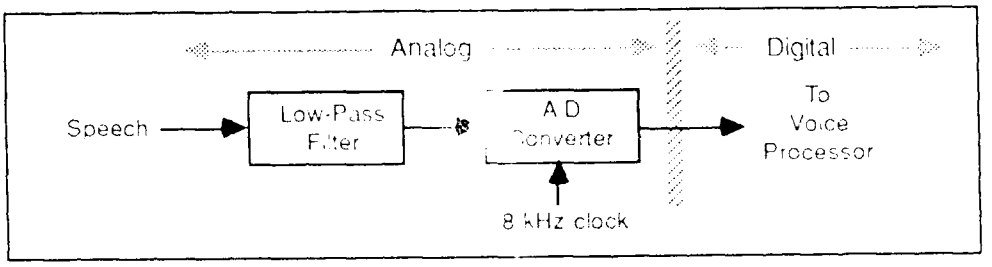
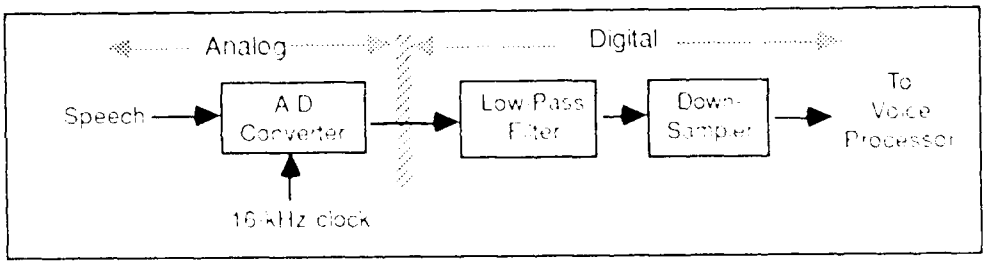


Fig. 14 — Speech spectrum of speech received by an HF channel at a distance of 500 mi from the transmitter. The AGC algorithm provided a perfectly matched speech level for ANDVT. This is a remarkably clear spectrogram of speech encoded at 2400 b/s and transmitted over a live HF channel. The presence of sharp speech onsets and clear formant structure implies that speech has been properly amplified by the AGC mechanism discussed in this section.



(a) Conventional approach



(b) Our approach

Fig. 15 — Old and new approaches to A/D conversion. In our approach, the necessary filtering is effected by digital computations, which can more readily attain ideal filtering characteristics, such as a sharp cutoff rate and a linear phase response over the entire passband, than can an analog filter (see Fig. 16).

4.1 Digital Antialiasing Filter

A typical antialiasing filter has a cutoff characteristic on the order of −180 dB per octave. For example, a 4 kHz antialiasing filter has roll-off characteristics of: 0 dB at 3.6 kHz, −20 dB at 4 kHz, −40 dB at 4.4 kHz, . . . , −180 dB at 7.2 kHz. The impulse response of an antialiasing filter may be obtained from the following Hamming-windowed Fourier series:

$$h(i) = \begin{cases} G\left[0.54 - 0.46\cos\left(\dfrac{2\pi i}{I}\right)\right]\left[0.5 + \displaystyle\sum_{n=1}^{N}\cos\left(\dfrac{n\pi i}{I} - 0.5\right)\right], & \text{for } 0 \le i \le I \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

where the factor G makes the sum of the impulse response unity (i.e., a DC gain of unity). The quantity I is the total number of impulse response samples and is related to the attenuation rate beyond the cutoff frequency. The quantity N is related to the cutoff frequency for a given value of I. The impulse response is symmetric with respect to the midpoint. Thus the phase response is linear.

A 4 kHz low-pass filter with a frequency roll-off rate of approximately −180 dB per octave may be realized by letting $I = 43$ and $N = 22$ in Eq. (12). On the other hand, a 6 kHz low-pass filter with a similar frequency roll-off characteristic may be realized by letting $I = 43$ and $N = 33$. The impulse responses of these filters are listed in Table 2, and their frequency responses are shown in Fig. 16.

4.2 Downsampler

If a 4 kHz antialiasing filter is used, the filter output is down-sampled by a factor of two to one. Thus every other sample is skipped.

On the other hand, if a 6 kHz antialiasing filter is used, the filter output is down-sampled by a factor of four to three. In other words, every four consecutive samples produces three consecutive samples. These three consecutive samples are obtained by interpolation. Thus

$$\begin{aligned} y(1) &= x(1) + (1/3)[x(2) - x(1)] \\ y(2) &= x(2) + (2/3)[x(3) - x(2)] \\ y(3) &= x(4), \end{aligned} \tag{13}$$

where $x(1)$, $x(2)$, $x(3)$, and $x(4)$ are four consecutive input samples, and $y(1)$, $y(2)$, and $y(3)$ are three consecutive output samples from the downsampler.

5. Speech Signal Conditioning

Four types of input interferences are often encountered in the speech waveform. They are an unintentional DC bias generated by the A/D converter, 60 Hz hum, digital noise picked up by the analog circuit, and ambient acoustic noise. We discuss methods for suppressing these interferences.

Table 2 — Impulse Responses of 4 and 6 kHz Antialiasing Filters. The 4 kHz filter is for speech encoding; the 6 kHz filter is for speech or speaker recognition.

Index (j)	Impulse Response $h(j)$	
	4 kHz cutoff $I = 43, N = 22$	6 kHz cutoff $I = 43, N = 33$
1 and 43	0.00103	−0.00005
2 and 42	0.00112	−0.00112
3 and 41	−0.00171	0.00219
4 and 40	−0.00174	−0.00232
5 and 39	0.00314	0.00068
6 and 38	0.00271	0.00271
7 and 37	−0.00557	−0.00629
8 and 36	−0.00396	0.00727
9 and 35	0.00930	−0.00329
10 and 34	0.00538	−0.00540
11 and 33	−0.01459	0.01482
12 and 32	−0.00687	−0.01846
13 and 31	0.02282	0.01079
14 and 30	0.00829	0.00831
15 and 29	−0.03505	−0.03119
16 and 28	−0.00954	0.04397
17 and 27	0.05581	−0.03242
18 and 26	0.01052	−0.01054
19 and 25	−0.10113	0.07938
20 and 24	−0.01113	−0.15601
21 and 23	0.31613	0.21606
22	0.51046	0.76181

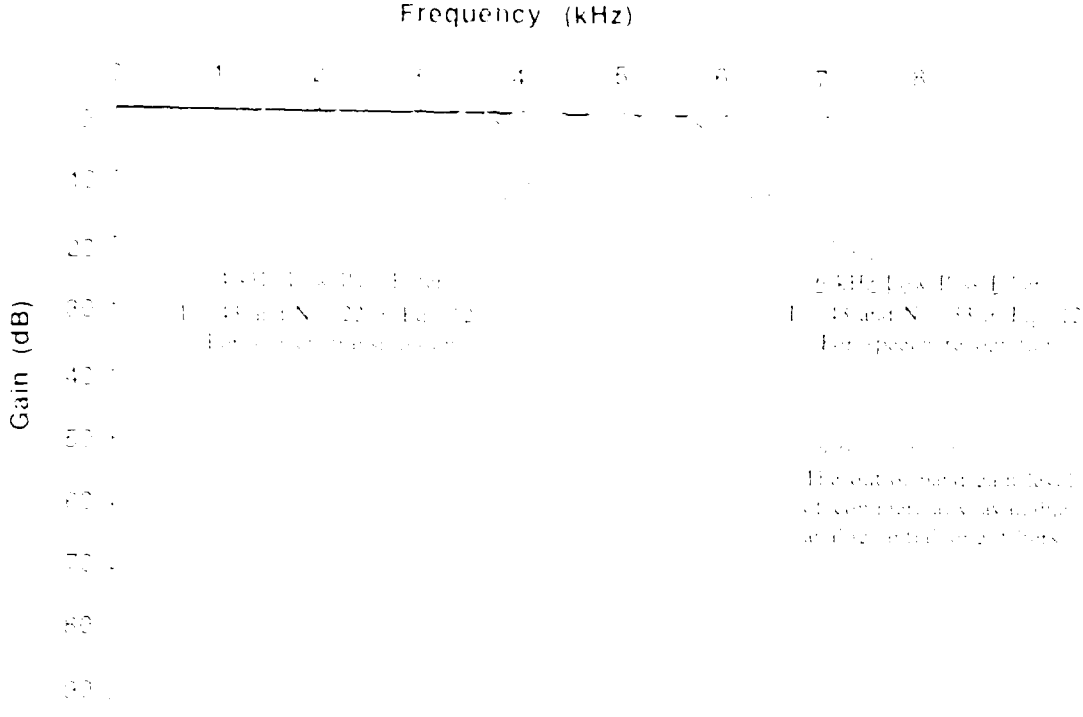


Fig. 16. Frequency response of 4 and 6 kHz antialiasing filters realized by digital filtering. Advantages of using digital antialiasing filters are: (1) in band frequency ripples are negligibly small (less than 0.03 dB), (2) frequency roll-off rates are steep (steeper than −180 dB/octave), (3) there are no return gains such as is often observed in analog filters, and (4) phase responses are linear functions of frequency (i.e., differential group delay is zero).

5.1 DC Bias Removal

DC bias is often generated within the A/D converter because of component deterioration in the output balance circuit. Table 3 lists the A/D converted output of our recently acquired signal processor. The magnitude of DC bias is alarmingly large. We cannot ignore the DC offset in the 12-bit A/D converter when its magnitude is as large as four bits (Table 3(a)). DC bias is probably generated after equipment has been deployed and people notice the degraded voice processor performance. Thus the voice preprocessor must be capable of removing DC bias.

In the LPC analysis, a speech sample is represented by a weighted sum of past samples (see Eq. (3)). In matrix notation, Eq. (3) may be represented by

$$X' = XA + E \tag{14}$$

The solution for A (prediction coefficients) that minimizes the mean-square errors is

$$A = (X^t X)^{-1} (X^t X'), \tag{15}$$

Table 3 — 12-Bit A/D Converter Output Samples (with the input grounded), (a) without DC suppression, (b) with DC suppression

−19	−19	19	−19	−19	−19	19	−19	0	0	0	0	0	0	0	0
−19	−19	−19	−19	−19	−19	−19	−19	0	0	0	0	0	0	0	0
−19	−19	−19	−19	19	−19	−19	19	0	0	0	0	0	0	0	0
−19	−19	−19	19	−19	−19	−19	19	0	0	0	0	0	0	0	0
−19	−19	19	−19	−19	−19	−19	−19	0	0	0	0	0	0	0	0
−19	−19	−19	−19	−19	−19	19	−19	1	1	1	1	1	1	1	1
−19	−19	−19	−19	−19	−19	−19	−19	0	0	0	0	0	0	0	0
−19	−19	−19	−19	−19	−19	−19	−19	0	0	0	0	0	0	0	0
19	−19	−19	−19	−19	−19	−19	−19	0	0	0	0	0	0	0	0
−19	−19	−19	−19	−19	−19	−19	−19	0	0	0	0	0	0	0	0
−19	−19	−19	−19	−19	−19	−19	−19	0	0	0	0	0	0	0	0

(a) (b)

where $(X^T X)$ is the autocorrelation matrix of the input speech samples. When speech amplitude is low and the DC bias is large, the autocorrelation matrix in Eq. (15) tends to have row elements with similar numerical values, making inversion of the matrix impossible. (The situation is similar to finding intersections of parallel lines.) Such an event creates a number of undesirable effects:

- The initial consonant intelligibility is reduced, particularly for /b/, /d/, and /n/, which are difficult to characterize even when DC bias is absent.
- The speech synthesizer tends to generate annoying pops or flutters when speech is absent.

The DC offset present in the A/D converter output may be removed by a simple DC-suppression filter made of a pole at $z = 1$ and a zero at $z = \alpha$, where $\alpha < 1$ (Fig. 17). Thus the transfer function of the DC suppression filter is expressed by

$$H_1(z) = \frac{1 + \alpha}{2} \frac{1 - z^{-1}}{1 - \alpha z^{-1}}, \tag{16}$$

where factor α is related to the 3 dB cutoff frequency (Table 4). The factor $(1 + \alpha)/2$ in Eq. (16) is to make the passband gain unity (i.e., $H_1(z) = 1$ at $z = -1$). Since this factor is nearly unity, it may be replaced with 1.0 to save the computation time; the consequence is that the speech amplitude becomes a fraction of a decibel lower. Figure 18 shows the frequency response of this DC-suppression filter.

5.2 60 Hz Hum Reduction

Often, faulty input coupling causes 60 Hz hum pickup. The presence of 60 Hz noise could be a serious problem for the digital voice processor. The design of a DC suppression filter is similar to the DC suppression presented in the preceding section. The zero and pole of the 60 Hz suppression filter have an argument that corresponds to ±60 Hz (i.e., $\pm\pi(60/4000) = \pm 0.015\pi$ radians), as illustrated in Fig. 19. Thus the transfer function of the 60 Hz suppression filter is

$$H_2(z) = G \frac{[1 - \alpha(e^{-j0.015\pi})z^{-1}][1 - \alpha(e^{j0.015\pi})z^{-1}]}{[1 - \alpha(e^{-j0.015\pi})z^{-1}][1 - \alpha(e^{j0.015\pi})z^{-1}]}, \tag{17}$$

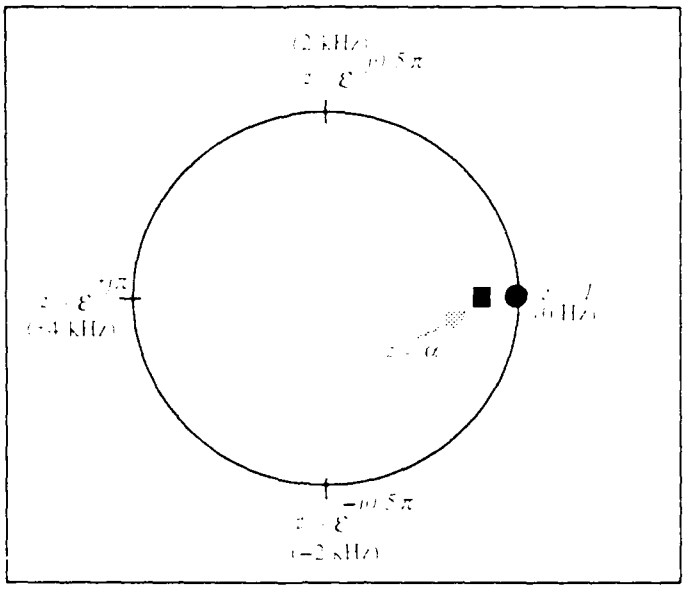


Fig. 17 — Zero and pole of the DC suppression filter. The zero located at $z = 1$ is indicated by ●; the pole located at $z = \alpha$ ($\alpha < 1$) is indicated by ■. The magnitude of α controls the cutoff frequency (see Table 4). Compare this figure with Fig. 19 (60 Hz suppression filter).

Table 4 — Cutoff Frequency in Terms of Filter Parameter α. Any α values between 0.875 and 0.925 are acceptable

Filter Parameter α	−3 dB Cutoff Frequency
0.855	200 Hz
0.865	185
0.875	170
0.885	156
0.895	142
0.905	128
0.915	114
0.925	100

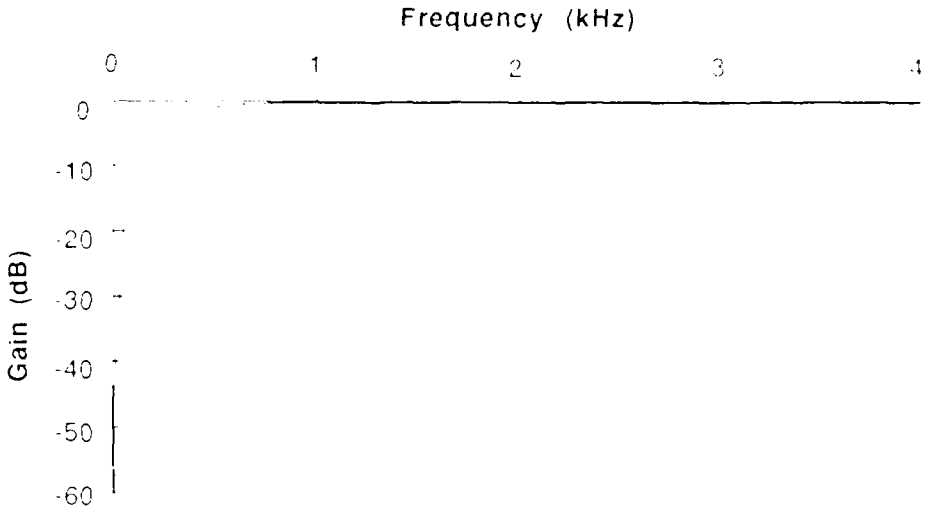


Fig. 18 — Frequency response of the DC suppressor. The response rises smoothly from 0 Hz, and no in-band frequency ripples occur that could be detrimental to the LPC analysis. This figure is plotted for $\alpha = 0.885$, and the -3 dB cutoff frequency is 156 Hz (see Table 4).

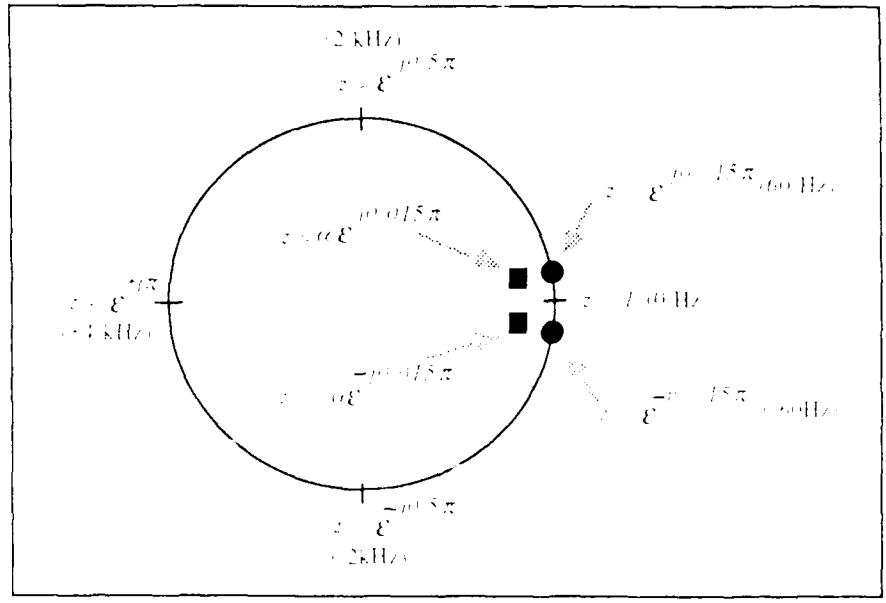


Fig. 19 — Zero and pole of the 60 Hz suppression filter. The zero is indicated by ●, the pole is indicated by ■. Compare this figure with Fig. 17 (DC suppression filter).

where α is the parameter that controls the notch bandwidth around 60 Hz, and G is a gain factor that makes the passband gain unity (i.e., $H_2(z) = 1$ at $z = -1$).

Equation (17) may be simplified as

$$H_2(z) = G\frac{1 - 1.99777988\,z^{-1} + z^{-2}}{1 - 2\alpha(0.9988899)z^{-1} + \alpha^2 z^{-2}}, \tag{18}$$

where

$$G = \frac{(1 + \alpha^2) + 2\alpha(0.9988899)}{3.99777988}. \tag{19}$$

Although the DC suppression filter has a wide range of acceptable values of α (see Table 4), the 60 Hz suppression filter has only a limited range of values of α because the frequency response must rise sharply beyond 60 Hz. We recommend $\alpha = 0.94$; and the transfer function of the 60 Hz suppression filter becomes

$$H_2(z) = 0.9409005\frac{1 - 1.99777988\,z^{-1} + z^{-2}}{1 - 1.877913\,z^{-1} + .8836\,z^{-2}}. \tag{20}$$

Figure 20 shows the frequency response of this 60 Hz suppression filter.

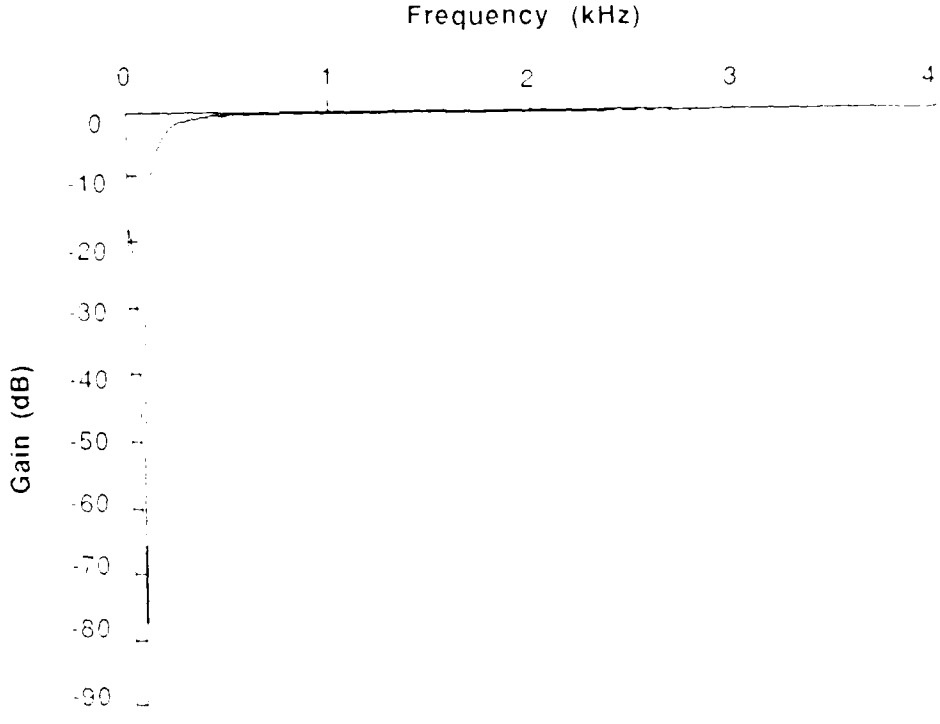


Fig. 20. Frequency response of a simple 60 Hz notch filter. Often, a faulty input coupling introduces 60 Hz noise. As noted, this 60 Hz suppression filter also removes the DC component by 17.23 dB. Thus this 60 Hz suppression filter can be used as a general-purpose, low-frequency cutoff filter.

5.3 Digital Noise Reduction

The presence of noise (such as digital noise pickup by the analog circuits) limits the available dynamic range. The digital noise pickup can be controlled by isolating the analog circuits from the digital circuits and by filtering power lines if a common power supply is used. We recommend that the magnitude of digital noise be within one or two least significant bits for the 12-bit A/D converter output. This criterion is not difficult to meet. We recommend that all analog circuits be placed in a separate copper can to minimize digital noise pickup.

5.4 Ambient Noise Reduction

When speech is transmitted digitally at low bit rates, speech intelligibility is degraded by as much as 15 to 25 points because of pitch and voicing errors and the inability of the filter coefficients to describe accurately the complex spectra of noisy speech. Likewise, a voice recognizer score that is 82% accurate at 85 dB sound pressure level (SPL) (i.e., office environments) scores only 13% in 115 dB SPL (i.e., helicopter platforms) [4]. Since voice processors often operate in noisy military platforms, reduction of ambient noise is a significant objective of preprocessing.

Spectral Noise Subtraction Method

We tested a spectral subtraction method that is a family of frequency-domain noise-reduction techniques; it subtracts the estimated short-term amplitude spectrum of noise from the short term amplitude spectrum of noisy speech. The resultant spectrum is converted to the speech signal by using the original input phase spectrum (implying that no steps are taken to refine the phase spectrum). This technique has been investigated extensively by Lim [5], Berouti et al. [6], Boll [7], Weiss et al. [8], and others. They all perceived that the output speech was improved by incorporating a number of artifacts, including the oversubtraction of spectra and the amplitude transformation of the individual spectrum. Thus the estimated speech spectrum is often denoted by the general form

$$|S(k)|^{\mu} = |Y(k)|^{\mu} - \gamma |N(k)|^{\mu}, \qquad k = 0, 1, 2, \ldots, 127, \tag{21}$$

where $|N(k)|$ is the estimated kth amplitude spectral component of noise that must be updated during the absence of speech, $|Y(k)|$ is the kth amplitude spectral component of the noise suppressor input; and $|S(k)|$ is the estimated kth amplitude spectral component of speech (i.e., noise suppressor output). In Eq. (21), $\mu \geq 1$ and $\beta \geq 1$. Berouti, Schwartz, and Makhoul [6] used $\mu = 2$ with an adjustable β; Boll [7] used $\gamma = 1$ and $\mu = 1$; and Weiss et al. [8] used, in effect, $\gamma = 1$ with an adjustable μ. We used a set of parameters (i.e., $\gamma = 1$ and $\mu = 2$). Thus,

$$|S(k)|^2 = |Y(k)|^2 - |N(k)|^2, \qquad k = 0, 1, 2, \ldots, 127. \tag{22}$$

Equalization of Microphone Response

We introduced equalization of the microphone frequency response in Eq. (22) by weighting the speech spectrum by the differential gain between the actual microphone response and the ideal flat response. The speech spectrum with equalized microphone response is expressed by

$$|S(k)|^2 = W(k)\,[\,|Y(k)|^2 - |N(k)|^2\,], \qquad k = 0, 1, 2, \ldots, 127 \tag{23}$$

where $W(k)$ is the kth differential gain (in power ratio, not in decibels) between the actual microphone response and the ideal flat response for the kth frequency. Note that the index k is incremented by a frequency step of (4000/128) = 31.25 Hz. We need not compensate the microphone response outside 150 and 3800 Hz. Thus, $W(k) = 1$ if $k < 5$ and $k > 122$.

Additional Factors

Even if the values of μ and γ are fixed (i.e., $\mu = 1$ and $\gamma = 2$), the noise-suppression performance is dependent on other salient factors not explicitly shown in Eq. (23). These factors are:

1. *Spectral analysis*—The 180 speech samples of the current frame were overlapped with the 76 trailing samples of the previous frame through trapezoidal windowing. We chose a frame of 76 samples because the resulting 256 samples permit the use of a standard FFT for the time-to-frequency transformation.

2. *Minimum spectral floor*—If the subtracted spectrum in Eq. (23) (i.e., the left-hand member) was less than zero, it was replaced with zero because the amplitude spectrum cannot be negative. But, as noted by Berouti, Schwartz, and Makhoul [8], a small amount of spectral floor improved the output speech quality. They used a minimum spectral floor of somewhere between −20 and −46 dB with respect to the estimated noise spectrum. In our DAM tests (where $\mu = 2$), we used a fixed value of −34 dB. Thus

$$|S(k)|^2 \geq 0.02|N(k)|^2, \quad k = 0, 1, 2, \ldots, 127. \tag{24}$$

3. *Noise spectrum updating period*—In the noise-suppression technique that makes use of a single microphone, the noise spectrum is available only when speech is absent. Therefore, the noise spectrum should be updated only in the absence of both voiced speech and unvoiced speech. In high-noise environments, however, unvoiced speech is difficult to detect because ambient noise is often much louder. Voiced speech has considerable energy at the first formant frequency, and most platform noises do not have strong resonant frequencies in the first formant region. As illustrated in Fig. 21, the histogram of low-band energy swings between larger values (when voiced speech is present) to smaller values (when voiced speech is absent). During each frame, we obtained the low-band energy of the current frame by simply summing the first 32 spectral power densities available for spectral subtraction. The past history of low-band energy was scanned to determine the maximum (MAX) and minimum (MIN) values. We updated the noise spectrum when the current low-band energy (P) was below a threshold set at one-eighth of the distance from MIN to MAX

$$P < \text{MIN} + (\text{MAX} - \text{MIN})/8. \tag{25}$$

Although we updated the noise spectrum during unvoiced frames, the effect was not too adverse because unvoiced speech was generally brief in comparison to the silent periods between phrases.

4. *Noise spectrum adaptation*—The first-order low-pass filtering given by

$$|N(k)|^2 \leftarrow G|N(k)|^2 + (1 - G)|Y(k)|^2, \quad k = 0, 1, 2, \ldots, 127 \tag{26}$$

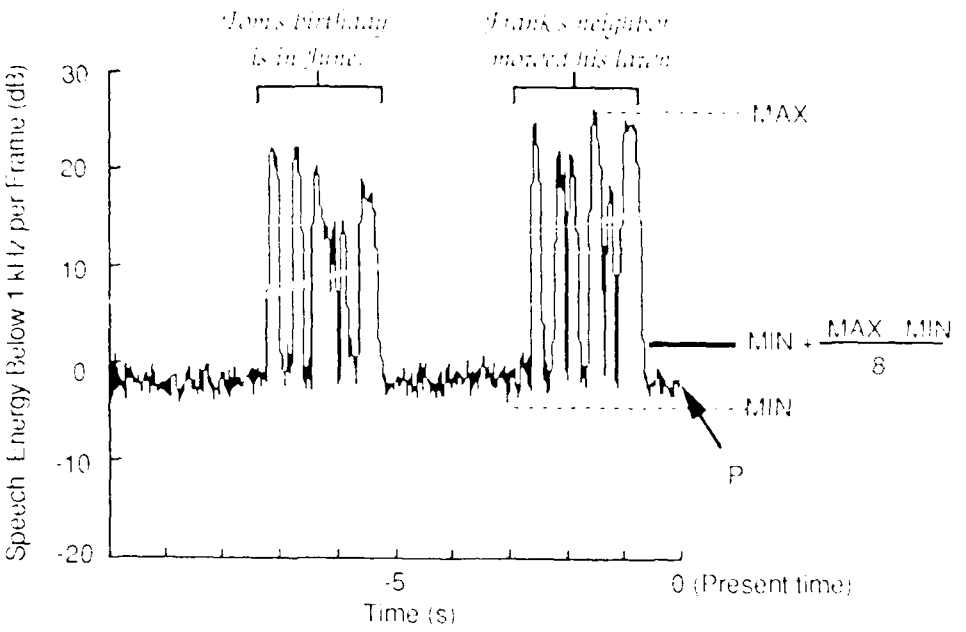


Fig. 21 — Histogram of speech energy below 1 kHz contained in the trapezoidally amplitude-weighted 256 samples (i.e., short-term averaged, low-band energy). The two sentences were spoken at a helicopter platform where the noise level was as high as 115 dB. When voiced speech is absent, the level of low-band energy is close to the minimum value (MIN) observed in the past 10-s history. The noise spectrum is updated when the low-band energy of the present frame (P) is below the threshold level that is indicated by the heavy line. As noted, the noise spectrum is updated during long gaps between sentences and brief gaps (a few frames) between words.

is adequate for updating the noise spectrum. In Eq. (26), G is a feedback factor that is normally $G = 15/16$. Quicker update is preferred when the input spectral density is less than the estimated noise spectral density. In this case, we used $G = 3/4$. This noise-spectrum adaptation method proved adequate. We observed that a suddenly appearing interfering tone during speech was effectively cancelled within a quarter of a second.

Prototype Performance

We selected 11 different types of noisy speech samples actually recorded at military platforms and an office. Figure 22 is an example of spectrograms before and after noise reduction when we used the speech samples recorded at a P3C cruising at a high altitude (the noise level is 105 dB). The noise suppressor reduced noise by 15 dB, and the output spectrum is remarkably free of ambient noise.

We also evaluated the prototype noise suppressor through the 2.4-kb/s speech encoder. The performance was evaluated by the standardized speech quality test, the Diagnostic Acceptability Measure (DAM). It evaluates the amount of speech distortion in terms of hissing, buzzing, rumbling, babbling, fluttering, muffled, nasal, unnatural, cracking, thin, and harsh. Hence the test would indicate the effectiveness of noise suppression when the voice processor is tested with and without noise suppression. The following is a summary of test results:

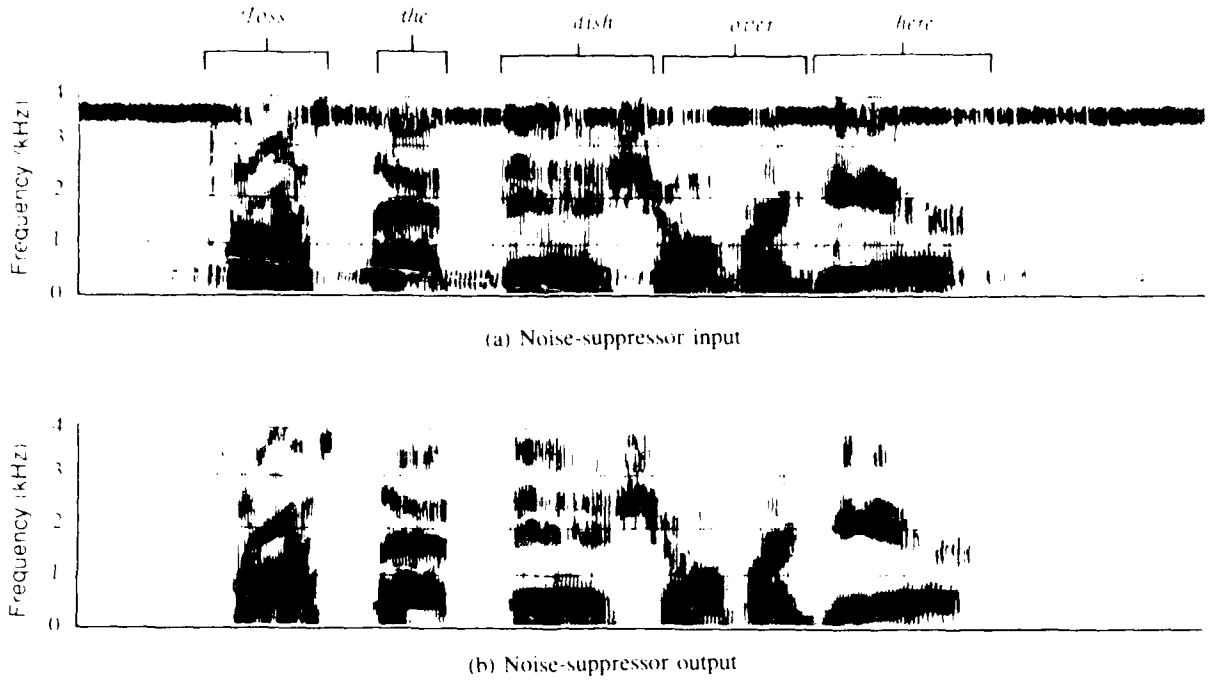


Fig. 22 — Spectrograms of noisy speech recorded at a P3C platform. The turboprop noise generated by the P3C is rather stationary with two prominent resonant frequencies. The noise suppressor removed most of the prop noise.

1. *Average quality score improvement*—The voice quality improved for every speech material we tested. The average score improved 6 points, which is substantial. In the shipboard environment where the noise level was only 76 dB, the score was improved by only 2.6 points (see Fig. 23).

2. *No adverse effects with noise-free speech*—Spectral subtraction did not degrade the quality of noise-free speech, contrary to some of the previously tested noise-suppression techniques. In fact, spectral subtraction removed even the background hiss produced by the original analog tape. As a result, speech quality improved from 50.2 to 52.7.

3. *Least improvement with nonstationary noise*—As expected, a noise-subtraction method making use of a single microphone did not perform well with nonstationary noise (such as the fluttering wind noise encountered in a moving jeep) because a slowly updated noise spectrum cannot compensate effectively for the rapidly changing incoming noise spectrum.

4. *Dramatic improvement with relatively stationary noise*—If the noise is relatively stationary as is the P3C noise, the performance of the spectral subtraction method is remarkable. The score improved from 32.0 to 45.2 , which is comparable to the quality improvement in a noise-free environment when random bit errors are reduced from 5% to 0.5%.

6. Spoken Voice

The spectral characterization of speech improves if the talker pronounces words slowly and distinctly. The use of a delayed sidetone is an effective way to induce the talker to articulate more

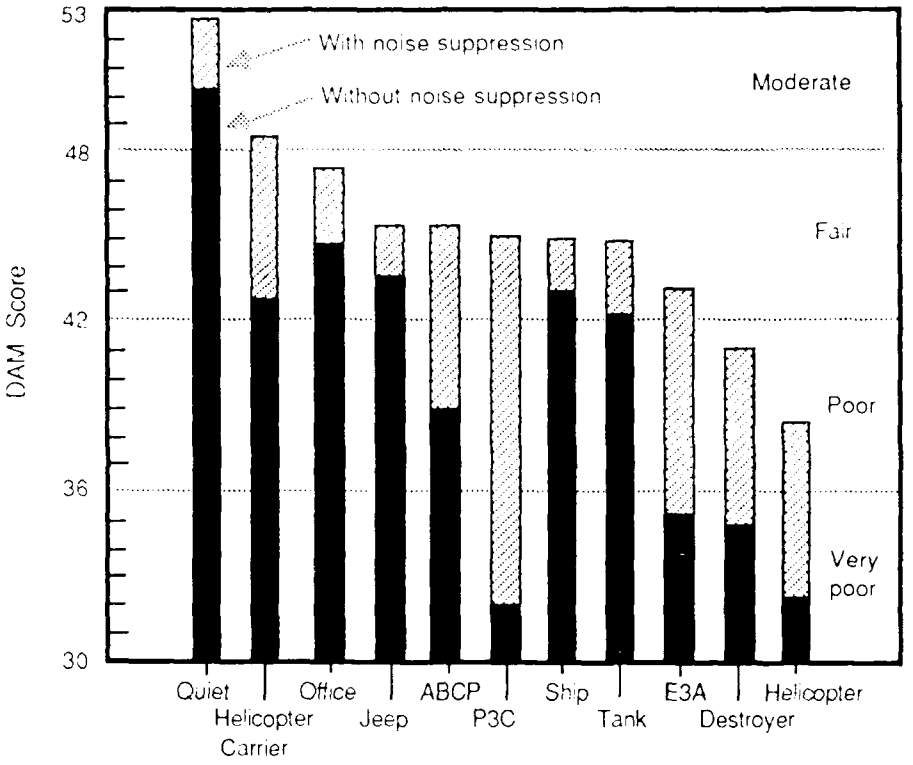


Fig. 23 — Speech quality scores with and without the spectral subtraction method. Speech quality was improved in all cases, and the speech quality was upgraded from "very poor" to "poor," "poor" to "fair," etc. These descriptive terms were devised by the Digital Voice Processor Consortium, which has been testing voice processors since 1972.

slowly and conscientiously. Likewise, if the speech spectrum is balanced between low and upper frequency bands, the result of speech analysis improves. We present a method of equalizing the speech spectral envelope.

6.1 Sidetone Considerations

Sidetone is an acoustic feedback of the speaker's own voice to the earphone of the handset used for transmission. In the full-duplex telephone, sidetone is superimposed with the received signal from the other end, including line noise and the speaker's voice.

Sidetone at the speaker's site performs many benefits. Richards [9] found that the relative loudness of sidetone influences how loud a person talks. The absence of any sidetone indicates to the talker that the line is dead. What is more important, the quality of sidetone gives the talkers some idea of the quality of the connection, which influences their manner of talking. For example, a very noisy line (as evidenced by a noisy sidetone) usually encourages the talker to speak louder. Likewise, a line with echo may influence the talker to speak more slowly and distinctly.

Black [10] found that delayed auditory feedback causes a slowdown in the talking rate in proportion to sidetone delays of up to 200 to 250 ms. Too much delay, however, causes articulation disturbances, which have been extensively studied and documented [11]. NRL also conducted delayed sidetone experiments that confirmed the previous findings; namely, talking slows down with increasing sidetone delays of up to 100 ms (Fig. 24) [12].

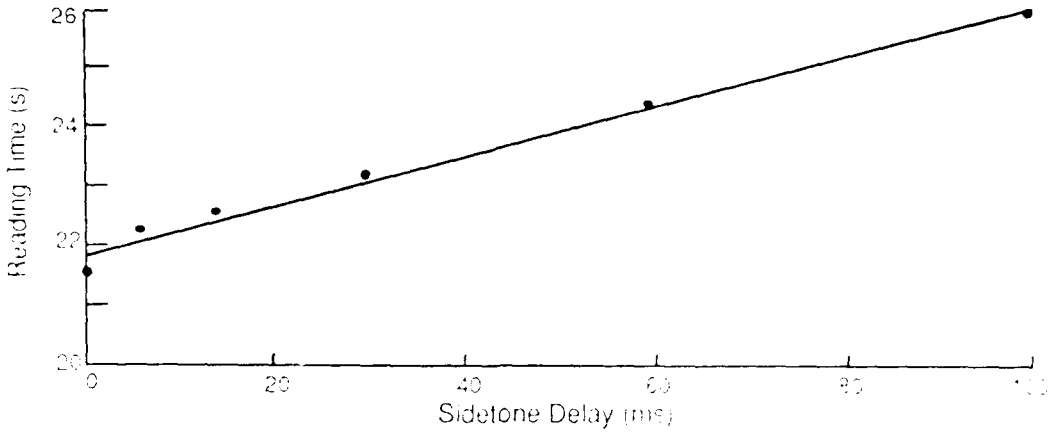


Fig. 24. Reading time vs sidetone delays. In this experiment, a list of 12 sentences and four 24-word lists (one list each of one, two, three, and four-syllable words) were read while hearing one's own voice delayed in the earphone of the handset [12]. The mean value of reading time with no delay is 21.6 s, and reading time increases linearly to 26 s with a delay of 100 ms.

Previously, NRL specified a delay of 30 ms for ANDVT, a tactical 2400-b/s secure telephone for tri-service use (mentioned in Section 3.2). A delay of 30 ms is relatively small, and it will not affect communication. (The Bell System does not use echo suppression if echo delay is 30 ms or less.) The usefulness of sidetone in actual environments will be further evaluated by the user reactions to ANDVT as they are deployed in quantity in the near future.

6.2 Speech Spectral Tilt Equalization

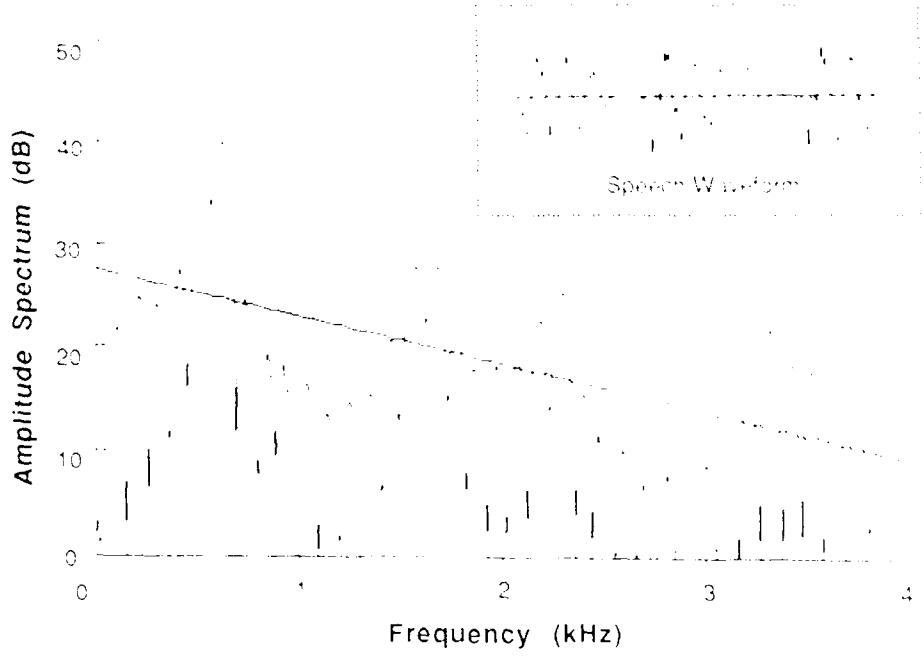
For a given speech sound, the amount of speech spectral tilt varies significantly from person to person (Fig. 25). A clear ringing voice has more high-frequency energies (Fig. 25(a)) because of the following favorable characteristics of the glottis and vocal tract: (a) glottis closes instantly (i.e., wide-band excitation), (b) glottis closes completely (i.e., a good "on-and-off" contrast), (c) vocal tract is not lossy (i.e., no speech leakage from the nasal passages). Other voices have weak upper bands (Fig. 25(b)) because their glottis and vocal characteristics are opposite of these. A speaker recognition process directly or indirectly exploits the spectral tilt to identify or verify speakers.

For other speech applications, however, a wide variation in the spectral tilt results in speaker-dependent performance because the LPC analysis does not work well with those speech signals having weak upper frequency components. Therefore, LPC analysis is often preceded by preemphasis (high-frequency boost). Usually, a fixed preemphasis is used. Since the magnitude of the spectral tilt is different from person to person, a preferred preemphasis would be an adaptive preemphasis whereby the amount of high-frequency boost is self-controlled by the amount of spectral tilt of the input speech.

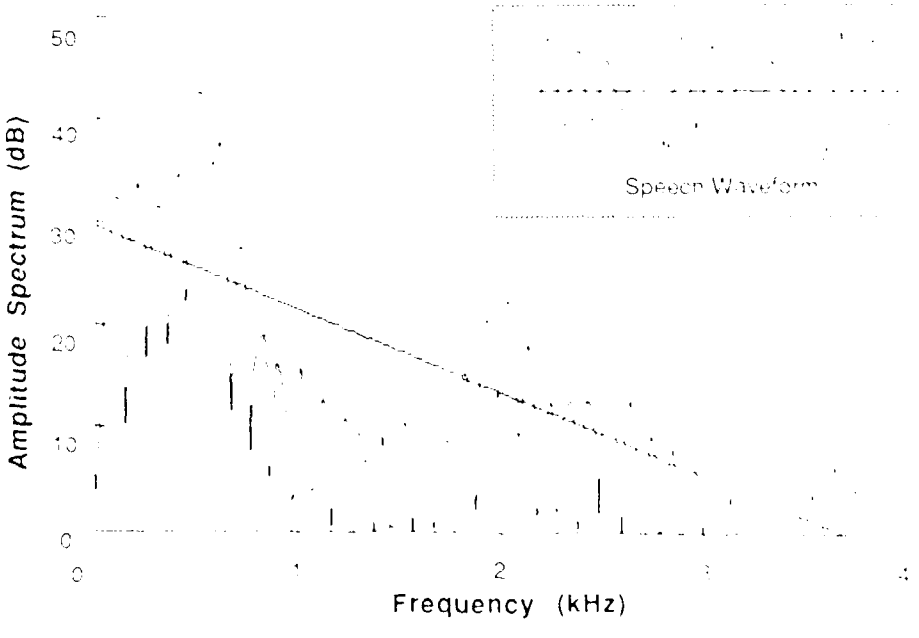
Adaptive preemphasis is accomplished by a single-zero filter with an adaptive filter weight:

$$y(i) = x(i) - \beta\, x(i-1), \tag{27}$$

where β is the adaptive preemphasis factor, and $x(i)$ and $y(i)$ are the input and output speech samples. We chose β to be the coefficient of the first-order linear predictor because it approximates the speech envelope by a single variable, and this variable contains mainly information regarding the spectral tilt. Thus,



(a) Speaker No. 1



(b) Speaker No. 2

Fig. 25 — Speech spectra of vowel "a" in "way" from two different talkers. (a) a clear and ringing voice that is not easily drowned by ambient noise (a good voice for cocktail parties). (b) a typical aging voice that lacks high-frequency energies. The LPC analysis disfavors the speech spectrum that is heavily tilted. Thus, LPC analysis is usually preceded by a preemphasis (high-frequency boosting); it has always been a fixed preemphasis.

$$\beta = \frac{E[x(i)x(i-1)]}{0.5\{E[x^2(i)] + E[x^2(i-1)]\}}, \quad -1.0 < \beta < 0.5 \tag{28}$$

where $E[\cdot]$ signifies the running average of the past history, which is on the order of 1 s. The theoretical range of β is -1.0 and 1.0. We, however, intentionally limit its lower range to 0.5 because if β is less than 0.5, the speech signal has strong high-frequency components (i.e., unvoiced fricatives /s/, /sh/, /ch/, etc.); hence, no further preemphasis is needed. Figure 26 is the frequency response of the adaptive preemphasizer for various values of preemphasis factors. Since the quantity β is derived through long-term averaging, it is more dependent on the speaker's vocal timbre than the spoken words (which has been smudged by long-term averaging).

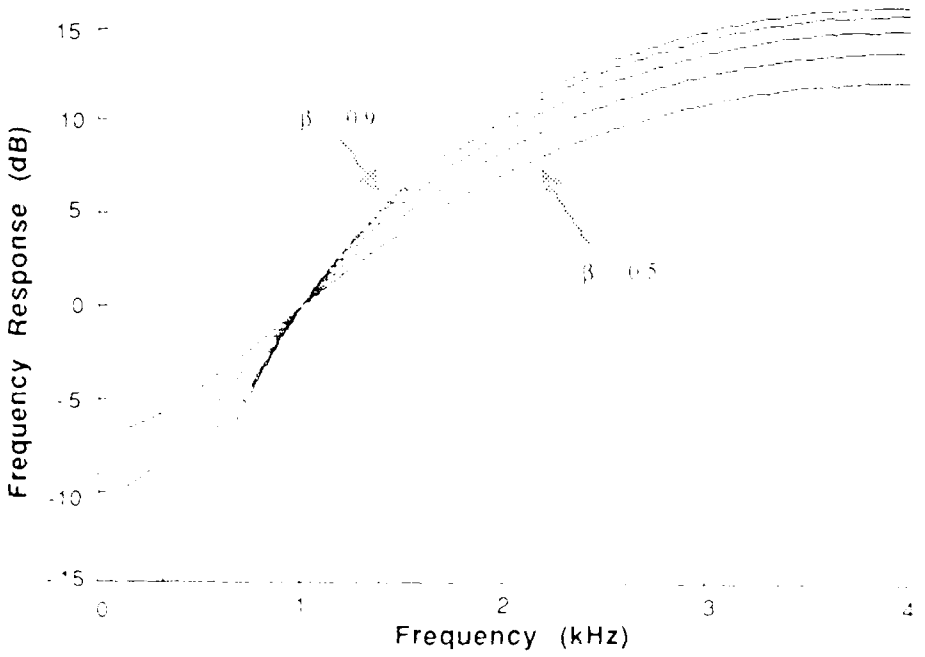


Fig. 26 — Adaptive preemphasis filter responses for various preemphasis factors (β) from 0.5 to 0.9 at an increment of 0.1. For comparison purposes, all the frequency responses are normalized to have a unity gain at 1000 Hz. A voice with strong high-frequency components (Fig. 25(a)) achieves a smaller β value of 0.774; therefore high frequencies are not boosted as much. On the other hand, a voice without strong high-frequency components (Fig. 25(b)) attains a large β value of 0.907; therefore high frequencies are boosted more than for the other voice. Thus both speech samples have more balanced lower and upper band spectral distributions.

CONCLUSIONS

In this report, we discuss how to preprocess the speech signal in such a way that the subsequent digital voice processing algorithm can function at its best. The approach presented here is applicable to speech coding, speech recognition, speaker recognition, or any other processor used to extract verbal or nonverbal information from speech.

A preprocessor is no longer a fixed-gain amplifier with an antialiasing filter. It is an adaptive system that can self-adjust the speech level and remove any interference (i.e., DC bias, 60 Hz hum, digital noise, ambient noise) if present. It can equalize a nonflat microphone response. The preprocessor also has an antialiasing filter that has an excellent roll-off characteristic (−180 dB per octave) with differential group delays of zero anywhere within the passband. The preprocessor even equalizes wide variations of speech spectral tilt to improve the quality of the extracted speech parameters. More importantly, the only analog circuit we use is a variable-gain amplifier at the front end and the A/D converter. Since no elaborate analog circuits are involved, the preprocessor is not a hindrance to hardware miniaturization.

This report is the result of our continuing effort to make voice processors more reliable and to operate successfully in difficult real-world environments.

ACKNOWLEDGMENTS

The authors thank Dr. John Davis, superintendent of the Information Technology Division at NRL, and we thank the Research Advisory Committee of NRL for supporting our research projects. We also thank Capt. Tucker, Robert Martin, Timothy McChesney, and Sharon James of SPAWAR for support of these development projects. Without their support, we could not have generated and tested the various ideas contained in this report.

REFERENCES

1. G.S. Kang and S.S. Everett, "Improvement of the Narrowband Linear Predictive Coder, Part 1—Analysis Improvements," NRL Report 8645, Dec. 1982.

2. G.S. Kang and S.S. Everett, "Improvement of the Narrowband Linear Predictive Coder, Part 2—Synthesis Improvements," NRL Report 8799, June 1984; "Improvement of the Excitation Source in the Narrow-Band Linear Predictive Vocoder," *IEEE Trans. Acoust., Speech, Signal Proc.* **ASSP-33**, 377-386 (1985).

3. H.F. Olson, "Gradient Microphones," *J. Acoust. Soc. Am.* **17**(3), 192-198 (1945).

4. C.H. Teacher and D.C. Coulter, "Performance of LPC Vocoders in a Noisy Environment," *IEEE*: **ICASSP**-79, 216-219 (1979).

5. J.S. Lim, "Enhancement and Bandwidth Compression of Noisy Speech by Estimation of Speech and Its Model Parameters," Sc.D. dissertation, Dept. of Elec. Eng. and Comput. Sci., MIT, Cambridge, Aug. 1978.

6. M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise," *IEEE ICASSP Record*, Apr. 1979, pp. 208-211.

7. S.F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. Acoust., Speech, Signal Proc.* **ASSP-27**, 113-120 (1979).

8. M.R. Weiss et al., "Processing Speech Signals to Attenuate Interference," presented at *IEEE Symp. Speech Recognition*, Apr. 1974.

9. D.L. Richards, *Telecommunications by Speech* (John Wiley and Sons, New York, 1973), p. 308.

10. J.W. Black, "The Reading of Messages of Different Types and Numbers of Syllables Under Conditions of Delayed Sidetone," *Language and Speech* **1**, 211-217 (1958).

11. R.A. Case et al., "Bibliography: Delayed Auditory Feedback," *J. Speech Hear. Res.* **2**, 193-200 (1967).

12. A. Schmidt-Nielsen and D.C. Coulter, "Effect of Modest Sidetone Delays in Modifying Talker Rates and Articulation in a Communication Task," *Proc. 97th Meeting of Acoust. Soc. Am.*, pp. 493-496 (1979).